Comparative Analysis and Explainability of Uplift Modeling Techniques for Targeted Marketing Campaigns

Pranav More
Department of Computer Science
University of Exeter
Exeter, UK
pm621@exeter.ac.uk

Dr. Aishwaryaprajna
Department of Computer Science
University of Exeter
Exeter, UK
aishwaryaprajna@exeter.ac.uk

***Abstract*—The project reviews and compares different uplift modeling techniques, developed with the aim to optimize targeted marketing campaigns by identifying customer segments most likely to react positively to certain interventions. Uplift modeling outperforms traditional predictive methods by predicting outcomes as well as estimating the causal impact of marketing activities. In this research, various uplift modeling techniques will be compared using the famous Hillstrom dataset with a real-world email marketing campaign. The model's effectiveness will be assessed in terms of uplift prediction accuracy, and the strengths and weaknesses of each technique in a marketing context will be discussed.**

Besides performance analysis, the study integrates explainable AI (XAI) techniques for model interpretability, specifically SHAP. This provides insight into factors that drive uplift and thus helps marketing strategists make informed, data-driven decisions. Results—Accented: The major findings indicated added value using uplift models instead of traditional approaches for targeted marketing by means of resource allocation and return on investment maximization.

***Index Terms*—Uplift Modeling, Explainability, Model Testing, Comparative Study**

I. Introduction

Targeted marketing today is a great way of doing business, compared with the past; it can maximize outreach to finally have a positive impact on the interests of customers. Among other things, being able to predict which customers respond well to which particular marketing actions is one of the most important aspects for maximizing return on investment within marketing activities. In this context, uplift modeling techniques are presented with regard to measuring incremental impact on a customer's behavior to provide a powerful analytical tool. Uplift is a term in the marketing domain and refers to the differences in the purchasing behaviour between customers who are offered the promotion (treated) and those who are not (control). [2] Uplift modeling aims at establishing the net difference in customer behavior that results from applying a specific treatment to customers, e.g., a reduction in the likelihood of churn when customers are targeted with the retention campaign. [1]

Nonetheless, the number of techniques in uplift modeling is rather enormous, and selecting the best one for the specific case is quite challenging. [3] Simultaneously, as much as accuracy, in the business world, transparency, and interpretability of these models is of great importance, given that the decisions have to be justified. This project addresses this dual challenge by comparing performance not just among different uplift modeling techniques—Single, Two-Model Classifier, T-Learner included—but also implementing Explainable AI methods, like SHAP, for unraveling the decision-making in these models.

What is innovative about this project is the holistic approach to the evaluation of efficacy related to the explanation of uplift models. By integrating XAI tools [4], this contributes to the ever-increasing interdisciplinary battleground of interpretable machine learning, which ultimately sheds relevant insights into the way better and transparent decisions are taken by businesses once adopted in their marketing strategies. The outcomes of the study will aid practitioners in deciding the right uplift model according to their needs and would further emphasize the importance of model interpretability in the larger context of data science and marketing analytics.

II. Related Work

Uplift modelling is a very niche technique in predictive analytics that measures the incremental impact of a treatment or action on individual behaviour—for example, a marketing campaign. Unlike the traditional classification model meant to predict the outcome using historical data, uplift models search for an estimate of the differential response between the treated and untreated groups. This differential insight comes in handy in targeted marketing, where one tries to find out who those customers are likely to be influenced by a certain marketing action.

Despite its practical appeal, uplift modelling has received surprisingly little attention in the literature. [5] Several methodologies have been developed for uplift modelling, each of them having their strengths and weaknesses. The most popular ensemble methods are bagging, boosting and Random

Forest. Other ensemble methods exist, such as Extremely Randomized Trees or Random Decision Trees. [5] Single-Learner methods incorporate treatment effects directly into the model for a unified prediction framework, often at the cost of interpretability. One study [2] mentions that the single-model approach is simple, easy to implement, and has the flexibility of being able to use any off-the-shelf supervised learning algorithm. However, a major drawback to this approach is that a single model may not model both potential outcomes well, and hence the estimation of CATE (condititional average treatment effect) may be biased. Another problem with the single-model approach is that treatment T may not be selected by a model that only uses a subset of the features for prediction (such as a tree model), and thus the CATE will be estimated as zero for all subjects

In contrast, the Two-Model Classifier [6] trains separate models for the treated and control groups at the cost of somewhat increased complexity and more straightforward interpretability. One study highlights the two-model approach to uplift modeling is based on the estimation of the two conditional expectations [11], $E[Y_i(1)|X_i]$ and $E[Y_i(0)|X_i]$, separately using the subsets of data. This method gains from its simplicity and flexibility. It is, in fact, possible to directly apply state-of-the-art machine learning algorithms like Random Forest and XGBoost. These models can be used in their standard forms, be it regression, binary classification, or multi-class classification, with no major changes needed in the uplift context.

In independently modeling treatment and control outcomes, this approach enables a practitioner to utilize the robust performance of these very well-established algorithms in guaranteeing an accurate prediction of the uplift effect. This also makes the modeling of complex interactions within each group easier: the treatment and control are divided, so it is easier to use advanced techniques in each. The method is popular in uplift modeling due to adaptability in a wide range of learning problems, thus typically balancing relatively easy implementation with high predictive accuracy.

Another strategy is the T-Learner, which models separately the treatment and control effects and calculates the uplift as the difference between these models. All those methods are performed based on machine learning techniques, mostly decision trees, random forests, and gradient boosting, hence attaining the complexities in uplift modelling. There is a study which positions the T-learner to be popular meta-algorithm [12] for estimating heterogeneous treatment effects. It works by separately estimating the expected outcomes for treated and control groups using base learners. The treatment effect is then calculated as the difference between these estimates. Commonly used with methods like linear regression and tree-based models, the T-learner offers flexibility in capturing variations in treatment effects across different subgroups, making it a valuable tool in causal inference and uplift modeling.

In one study [10], a specialized variant of Support Vector Machines (SVM) tailored for uplift modeling is presented. The SVM optimization problem is reformulated to explicitly capture the difference in class behavior between two datasets. The model predicts whether an individual will exhibit a positive, neutral, or negative response to an action. By adjusting a specific parameter, the model's sensitivity can be fine-tuned to control the proportion of neutral predictions. The dual coordinate descent method is adapted for efficient optimization, and the method is experimentally compared with other uplift modeling approaches, demonstrating its efficacy.

Explainable AI has become a quite important area of research, in particular in scenarios when model transparency is critical. Among the more popular methods in explainable AI is SHapley Additive exPlanations (SHAP) [8], which provides a consistent approach for the interpretation of model predictions through attribution of contribution by each feature to the final decision. In the case of uplift modelling, SHAP values will bring out how various attributes of the customer impact the model's prediction of uplift, hence making the recommendations more transparent and trustworthy.

Recent literature calls for a tradeoff between predictive accuracy and model interpretability in uplift modelling. In one study [9], a new metrics yielded remarkably higher profit across different uplift models, targeting depths, profit functions, and data sets. They further contribute to the growing field of interpretable data science by uncovering interdependencies between covariates, ITE(individual treatment effect), and profit and by clarifying whether customers are worth targeting because of high responsiveness or high value. Some studies have shown that while complex models—like gradient boosting [7]—may be able to offer improved predictive performance, their opacity can render them of very limited practical applicability, particularly in business environments where decision-makers need their insights to be clear and justifiable. The project has been motivated by these findings to apply SHAP in the evaluation of interpretability across various uplift models [8], going a level deep into an analysis that brings together performance with explainability. This shall be a prospective addition to the methodological development in uplift modelling and also help in meeting the ever-growing demand for more transparent AI systems in data-driven decision-making.

III. Aims and Objectives

The main aim of this project is to conduct an in-depth analysis and comparison of different uplift modelling techniques with respect to targeted marketing campaigns and back that with proper explanations. Specific objectives: The research will establish which one of the uplift modelling techniques—Single-Learner, Two-Model Classifier, or T-Learner—offers both accurate and interpretable predictions of the customer response to the different marketing strategies.

A. Objectives

1) *Development and Implementation of Uplift Models:* Implement three major uplift modelling techniques: Single-Learner, Two-Model Classifier, and T-Learner. These models will be trained and evaluated against

real-world data, which will create a benchmark for the performance of each technique. The Single-Learner approach consolidates treatment information in combination with other model features. Two-Model Classifier trains separate models for the treatment and control groups, and T-Learner applies two models for different treatments.

2) *Performance Evaluation:*
Performance will be evaluated by checking the predictive accuracy of these models using key metrics such as uplift at k, which is the increase in positive responses when targeting at some thresholds in the population. Other relevant performance metrics within this business setting are overall uplift and precision.

3) *Explainability and Interpretability Analysis*:
SHapley Additive exPlanations, or SHAP, helps measure the feature importance of each model, thereby giving insight into how customer attributes impact uplift prediction. To some extent, this is also important to understand the intrinsic drivers of the model's decisions and for results from the model to be trusted and interpretable by non-technical stakeholders.

4) *Comparative Analysis of Models:*
This will involve a comparative analysis of the models in terms of their trade-offs on predictive accuracy versus explainability. Or, more precisely, it will define which model among the considered ones is able to provide the best balance between very high accuracy of uplift prediction and easy interpretability in the process of decision-making.

These goals will, therefore, help in adding valuable information to this vital area of Data Science, especially in the application of machine learning models for targeted marketing. The findings will help organisations make sound decisions regarding which uplift modelling techniques to use depending on their particular needs and constraints.

IV. Experiment Design & Methods

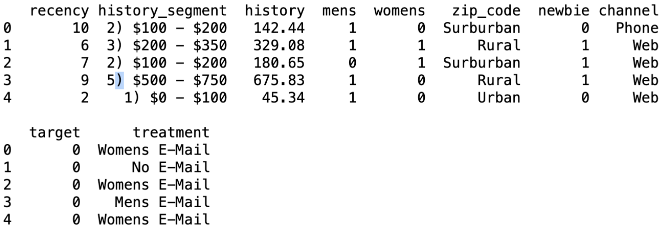
A. Overview of the Experimental Pipeline

The whole experimental process is thus divided into a few major stages: data collection and preparation, model selection and training, evaluation, and lastly, explainability analysis. Each phase is meticulously designed to guarantee the robustness and accuracy of the results, as well as the interpretability of the uplift models.

B. Dataset Description

The dataset used in this project is the Hillstrom dataset, which is commonly used for uplift modeling. The dataset originates from a real-world email marketing campaign conducted by Kevin Hillstrom. It contains 64,000 customers who were targeted with three different email campaigns: "Men's Email," "Women's Email," and a "No Email" control group.

1) Features: The dataset includes features such as customer demographics, previous purchasing history, and email engagement metrics. Specifically, it contains 12 features (Fig. 1.) like visit, conversion, and spend, which are the primary metrics of interest for understanding the impact of different marketing treatments. We primarily use the 'visit' feature as our target variable in our research as it provides the most engagement from a single customer/user as a reaction to any marketing campaign as compared to spend or conversion.



```
   recency history_segment  history  mens  womens   zip_code  newbie channel
0       10  2) $100 - $200   142.44     1       0  Surburban       0   Phone
1        6  3) $200 - $350   329.08     1       1      Rural       1     Web
2        7  2) $100 - $200   180.65     0       1  Surburban       1     Web
3        9  5) $500 - $750   675.83     1       0      Rural       1     Web
4        2    1) $0 - $100    45.34     1       0      Urban       0     Web

   target      treatment
0       0  Womens E-Mail
1       0      No E-Mail
2       0  Womens E-Mail
3       0    Mens E-Mail
4       0  Womens E-Mail
```

Fig. 1. Data Head - Features

C. Data Preparation

1) Missing Values: This is a very well-structured dataset. There are no missing values, which somehow makes analysis easier.
2) Categorical Variables: Categorical variables, like treatment—indicating the type of email a given subject received—went through one-hot encoding to be put into numerical values, thus ready for use by machine learning models.
3) Feature Engineering: New features were engineered from existing features to capture finer-grained customer behaviors. For instance, the visit feature was used to construct interaction terms with the treatment variable to understand conditional treatment effects on customer engagement.
4) Splitting the data: The data was split into a train-test set such that each subset is representative of the entire dataset with respect to treatments and outcomes.

D. Data Exploration

The data exploration phase provides valuable insights into the underlying structure and distribution of the dataset. This section will focus on three key aspects of the data: the target distribution, the treatment distribution, and the target distribution by treatment group. Each of these analyses helps to understand the dataset's characteristics and informs the subsequent modeling process.

1) Target Distribution: The target distribution reflects the proportion of positive and negative outcomes (here, visits) across the entire dataset. Understanding this distribution is crucial as it indicates whether the dataset is balanced or skewed. A balanced dataset is often preferable as it ensures that models do not become biased towards the majority class. In Fig. 2., the target distribution graph reveals whether there

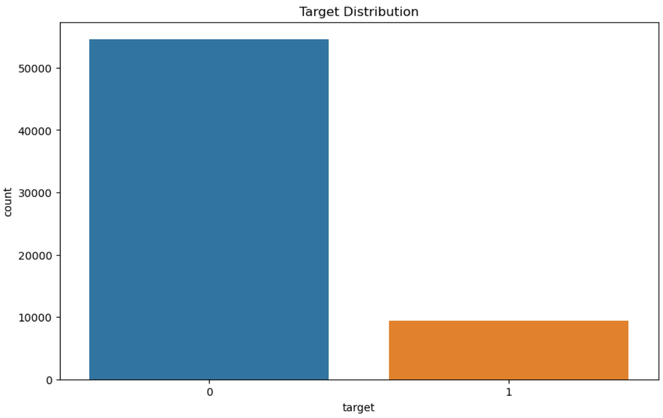


Fig. 2. Target Distribution Plot (Visit)

is an equal number of positive and negative responses, which directly impacts the model’s ability to accurately predict uplift.

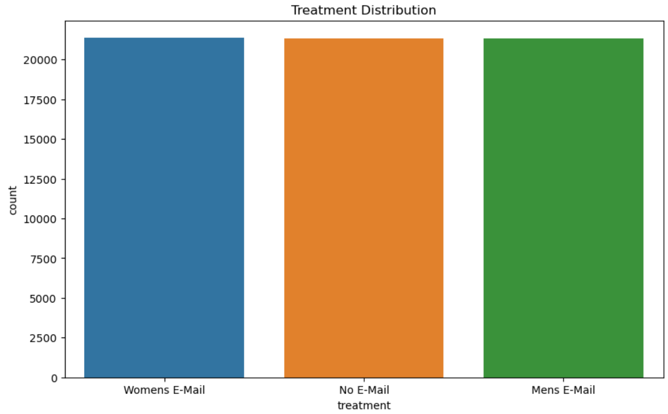


Fig. 3. Treatment Distribution

2) Treatment Distribution: The treatment distribution shows how the dataset is divided between the control group and the treatment group. This analysis is essential for understanding the impact of the treatment and ensuring that both groups are well-represented in the data. An imbalance in treatment distribution could lead to biased uplift estimates, as the model might overfit to the more represented group. The corresponding graph in Fig. 3. allows us to confirm that the treatment and control groups are adequately balanced.

3) Target Distribution by Treatment Group: Examining the target distribution by treatment group involves comparing the outcomes within the control and treatment groups separately. This analysis is crucial for uplift modeling as it directly relates to the model’s goal: predicting the differential impact of the treatment. The graph here in Fig. 4. helps to identify any significant differences in the target distribution between the two groups, which is a key factor in determining the potential uplift effect. By visualizing this distribution, we can assess the effectiveness of the treatment and guide the model training process.

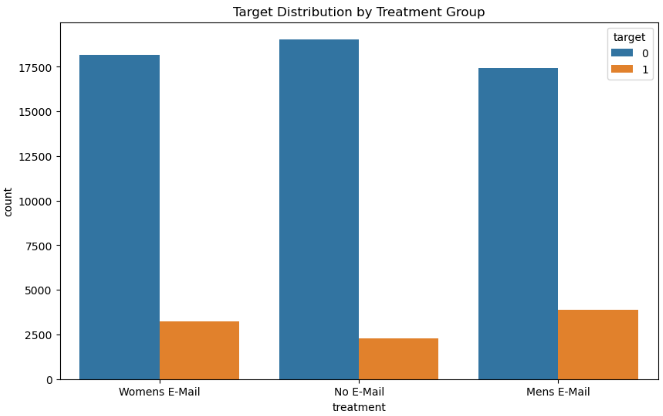


Fig. 4. Target Distribution by Treatment Group

E. Model Selection Criteria

1) Complexity and Flexibility: **Random Forest and XGBoost** were selected for their ability to capture non-linear relationships and interactions among features [13], which is crucial for accurately estimating treatment effects. These models are robust against overfitting due to their ensemble nature and perform well on structured data, such as the Hillstrom dataset used in this project.

2) Interpretability vs. Predictive Power: **T-Learner** and **Two-Model Approaches** were chosen for their interpretability and straightforward application in real-world scenarios. These methods separately estimate outcomes for the treatment and control groups [14], making it easier to understand the contribution of each feature to the uplift effect. This interpretability is essential in industries like marketing, where understanding the ”why” behind a model’s predictions is as important as the predictions themselves.

3) Scalability and Efficiency: Both **Random Forest** and **XGBoost** are scalable and efficient, making them suitable for large-scale applications [13] in real-world scenarios. Their ability to handle large datasets with high-dimensional features ensures that they can be deployed in production environments where quick decision-making is essential.

F. Model Training & Implementation

1) Single Learner (XGBoost): The Single Learner model was implemented using the XGBoost algorithm, which is known for its efficiency and performance in classification tasks. In this approach, the treatment indicator was included as an additional feature in the training data. The model was trained to predict the target directly, incorporating both features and treatment information simultaneously. This approach simplifies the uplift modeling process by handling both treatment and control groups within a single model [15]. XGBoost’s ability to handle complex interactions and its regularization techniques made it a suitable choice for this task. In the Single Learner model using XGBoost, the model’s objective is to minimize a loss function L that incorporates both the treatment T and the features X to predict the target y. Mathematically, the model can be expressed as:

$$\hat{y} = f(X, T; \theta)$$

where:

- $\hat{y}$ is the predicted target. - X represents the features. - T is the treatment indicator (0 for control, 1 for treated). - θ denotes the parameters of the model, which are learned during training.

The objective function for XGBoost is:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \Omega(f)$$

where:

- $L(y_i, \hat{y}_i)$ is the loss function, typically the logistic loss for binary classification. - $\Omega(f)$ is the regularization term that controls the complexity of the model to prevent overfitting.

The model then optimizes this objective function using a gradient boosting approach, iteratively improving the model by adding trees that correct the residuals of the previous trees. Each tree f_m is added to minimize the residuals:

$$f_m(X, T) = \arg\min \sum_{i=1}^{n} L(y_i, \hat{y}_{i,m-1} + f_m(X_i, T_i))$$

This approach allows XGBoost to efficiently handle the uplift modeling task by directly incorporating the treatment effect into the model's predictions.

2) Custom Two-Model: The Custom Two-Model approach involved training two separate models: one on the treatment group and another on the control group. These models were then used to predict outcomes for their respective groups. The difference between these predictions provided the uplift score. Random Forest classifiers were chosen for both models due to their robustness and ability to handle large datasets with many features. This method allowed for more targeted predictions within each group, ensuring that the model could capture specific effects within the treatment and control groups independently [16]. For the Custom Two-Model approach, the mathematical formulation involves training two separate models:

- Treatment Model $f_1(X)$: Trained on data where the treatment $T = 1$.
- Control Model $f_0(X)$: Trained on data where the treatment $T = 0$.

The uplift score $U(X)$ is then calculated as the difference between the predictions of these two models:

$$U(X) = f_1(X) - f_0(X)$$

3) SKLift RF Two-Model: The SKLift RF Two-Model was implemented using the 'TwoModels' class from the SKLift library. This approach is similar to the Custom Two-Model method but leverages the built-in capabilities of the SKLift library for uplift modeling. Two Random Forest classifiers were trained separately on the treatment and control groups. The uplift score was calculated by subtracting the control group predictions from the treatment group predictions. This approach benefits from the simplicity and efficiency of the Random Forest algorithm while taking advantage of the specialized uplift modeling functionalities provided by SKLift.

4) T-Learner (Gradient Boosting Method): The T-Learner approach involved training two separate Gradient Boosting classifiers: one for the treatment group and one for the control group. This method first estimates the conditional expectations of the outcomes for both treated and control groups separately. The difference between these estimates gives the uplift score. Gradient Boosting was chosen for its ability to model complex relationships and its effectiveness in handling imbalanced datasets, making it well-suited for this approach where accurate estimation of treatment effects is crucial.The mathematical formulation of the T-Learner can be expressed as follows:

1. Model Training:

- Train the treatment model $f_1(X)$ to estimate the expected outcome for the treated group:

$$\hat{Y}_1(X) = \mathbb{E}[Y|X, T = 1]$$

- Train the control model $f_0(X)$ to estimate the expected outcome for the control group:

$$\hat{Y}_0(X) = \mathbb{E}[Y|X, T = 0]$$

2. Uplift Calculation: - The uplift score for a given instance X is calculated by taking the difference between the predictions of the two models:

$$U(X) = \hat{Y}_1(X) - \hat{Y}_0(X)$$

This score represents the estimated effect of the treatment on the outcome, with positive values indicating a beneficial treatment effect, and negative values suggesting the treatment might be harmful. The Gradient Boosting classifiers were chosen for their capacity to model non-linear relationships and handle complex interactions between features, making them ideal for estimating these conditional expectations [17].

G. Deriving Explanations

To enhance the interpretability and transparency of the models, explainability techniques were employed, focusing on SHAP (SHapley Additive exPlanations) summary and dependency plots, as well as feature importance plots. These methods provide insights into how individual features contribute to model predictions, helping to understand the decision-making process behind each model.

1) SHAP Summary Plots: The SHAP summary plots were generated for both the Single Learner (XGBoost) and T-Learner models to provide insights into feature importance. For the Single Learner model, the SHAP plot highlighted the significant impact of features like recency and history on the model's predictions. Similarly, in the T-Learner approach, SHAP plots were created for both treated and control models, revealing how the influence of features, particularly recency, varied between treatment conditions.

2) SHAP Dependency Plots: Dependency plots were created for the Single Learner model to explore the interaction between specific features and their influence on model predictions. For demonstration purposes, a dependency plot for the ’recency’ feature was generated. An option to change this feature has been given within the code. This plot shows how the model’s prediction changes as the value of the recency feature varies, while also considering the interaction with another feature, such as ’history’. This provides a deeper understanding of the feature’s impact on the model’s decision-making process.

3) Feature Importance Plots: Feature importance plots were generated for all the models to understand the influence of different features on their predictions.

For the ”Single Learner model (XGBoost)”, the feature importance was directly extracted from the model, highlighting key features such as recency and history that the model relied on heavily.

In the ”Custom Two-Model” approach, feature importance was analyzed separately for the treatment and control models using Random Forest classifiers. The plots provided insights into which features were critical for the predictions in each group.

Similarly, the ”SKLift RF Two-Model” method involved generating feature importance plots for both treatment and control models, offering a comparison to the Custom Two-Model approach.

Lastly, the ”T-Learner” approach used Gradient Boosting classifiers to create feature importance plots for both treated and control models. These plots revealed how features influenced the model’s understanding of treatment effects under different conditions, highlighting the differential impact of features like recency across models.

These explainability techniques not only enhanced the transparency of the models but also provided actionable insights that could be leveraged to refine marketing strategies [18] based on the models’ predictions. By understanding the key drivers of uplift, decision-makers can better target their interventions, leading to more effective marketing campaigns.

H. Comparison Plots

To evaluate and compare the performance of the models, uplift curve line graphs were generated for each model, providing a visual representation of the cumulative uplift across different segments of the population. This allows for an intuitive comparison of how well each model identifies and ranks individuals most likely to respond positively to a treatment. Additionally, accuracy plots were created to further assess and compare the overall predictive performance of the models, highlighting their effectiveness in real-world scenarios. These visual tools were instrumental in identifying the strengths and weaknesses of each model.

V. Results

In this section we will a comprehensive overview of the outcomes from the various uplift modeling techniques applied in this study. It includes a detailed analysis of model performance through visualizations such as uplift curves, feature importance plots, and SHAP summaries. Each model’s ability to predict the incremental impact of marketing campaigns is evaluated and interpreted, highlighting key insights into their effectiveness and the importance of specific features in driving predictions.

A. Model Output Report

An uplift score indicates how much more likely a particular group is to take the desired action (e.g., making a purchase) when exposed to a treatment (e.g., receiving an email) compared to not receiving the treatment. For example, an uplift score of 0.05 means that the treatment increases the likelihood of the desired action by 5% compared to no treatment. [22]

1) Single-Learner XGBoost:

- Men’s Campaign: The model achieved an uplift score of 0.0580, indicating that the treatment increased the likelihood of a positive response by 5.8% compared to the control group.
- Women’s Campaign: The uplift score for women was 0.0377, showing a 3.77% increase in response due to the treatment.

2) SKLift TwoModels Random Forest:

- Men’s Campaign: This model reported an uplift score of 0.0624, suggesting that the treatment group was 6.24% more likely to respond positively than the control group.
- Women’s Campaign: The uplift score was lower at 0.0162, indicating a 1.62% increase in positive response for the treatment group.

3) Custom Two-Model Classifier:

- Men’s Campaign: The uplift score was 0.0647, reflecting a 6.47% increase in the likelihood of a positive response for the treated group.
- Women’s Campaign: The uplift score was 0.0187, indicating an uplift of 1.87% in the treatment group.

4) T-Learner:

- Men’s Campaign: This model yielded an uplift score of 0.0639, suggesting a 6.39% increase in response likelihood due to treatment.
- Women’s Campaign: The score for women was 0.0306, showing a 3.06% uplift in response due to treatment.

Model	Men’s Uplift Score	Women’s Uplift Score
Single-Learner XGBoost	0.0580	0.0377
SKLift TwoModels Random Forest	0.0624	0.0162
Custom Two-Model Classifier	0.0647	0.0187
T-Learner	0.0639	0.0306

TABLE I
Summary Table of Uplift Scores

Table 1. and the explanations help illustrate the performance of each model in increasing the likelihood of positive outcomes for both the men’s and women’s campaigns.

B. *Visual Analysis of Uplift Curves Across Models*

An analysis of the performance of various uplift models using uplift curves has also been done, which provides a visual representation of each model's effectiveness in identifying and targeting the most responsive segments of a population. Below, we discuss two models as examples, but similar analyses were conducted for all models explored in this study.

Single-Learner XGBoost vs. Custom Two-Model Classifier

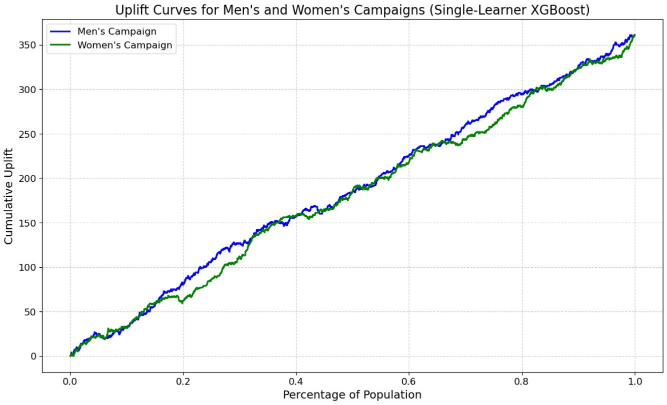


Fig. 5. Uplift Curve for Single Learner

1) *Single-Learner XGBoost:*

- The uplift curve for the Single-Learner XGBoost (Fig. 5.) model demonstrates a consistent rise in cumulative uplift across the population for both the men's and women's campaigns.
- The men's campaign shows a marginally higher uplift, suggesting the model is slightly more effective at identifying responsive segments within the male population.
- The steady slope of the curve indicates that the model efficiently ranks the population, consistently identifying those most likely to respond positively to the campaign.

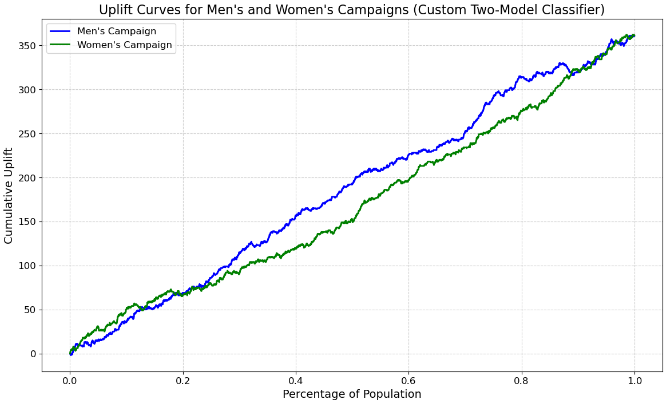


Fig. 6. Uplift Curve for Custom Two-Model

2) *Custom Two-Model Classifier:*

- The Custom Two-Model Classifier displays a more pronounced uplift(Fig. 6.), particularly towards the end of the curve, which indicates its capability to better distinguish between highly and less responsive individuals.
- The uplift curve for the men's campaign is consistently higher than that of the women's campaign, similar to the Single-Learner model, but with a slightly steeper and smoother increase.
- This suggests that while both models are effective, the Custom Two-Model Classifier may offer a slight advantage in maximizing response rates, especially in scenarios where precise targeting is critical.

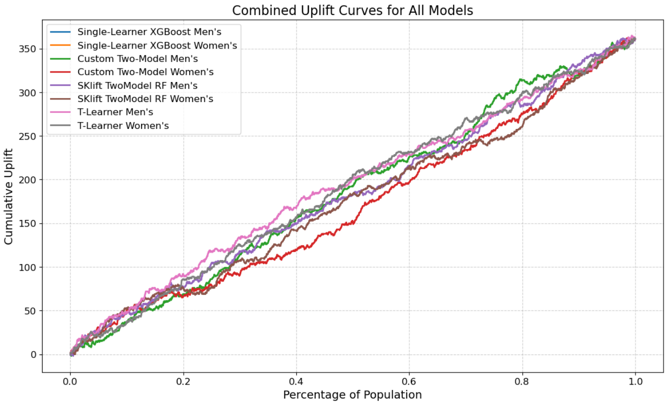


Fig. 7. All Models Uplift Curves

C. *Combined Uplift Curves for All Models*

Fig. 7. presents the combined uplift curves for all the models. This plot illustrates the cumulative uplift achieved by each model across different segments of the population. The uplift curves are crucial for understanding how effectively each model distinguishes between the treatment and control groups and thus optimizes the treatment effect.

- Single-Learner XGBoost (Men's and Women's) models demonstrate moderate performance throughout the population, with steady increases in cumulative uplift.
- Custom Two-Model (Men's and Women's) models show higher initial uplift, especially in the early segments of the population, suggesting that these models are effective at identifying segments with significant treatment effects early on.
- SKlift TwoModel RF (Men's and Women's) models exhibit a consistent uplift, with a slight advantage for the men's model across most of the population.
- T-Learner (Men's and Women's) models, while generally comparable, show a lower overall uplift compared to other models, indicating less effective differentiation between treatment and control groups.

D. *Model Accuracy Comparison*

Fig. 8 provides a comparison of the model accuracies. This plot is essential to assess not only the uplift but also the predictive accuracy of the models.

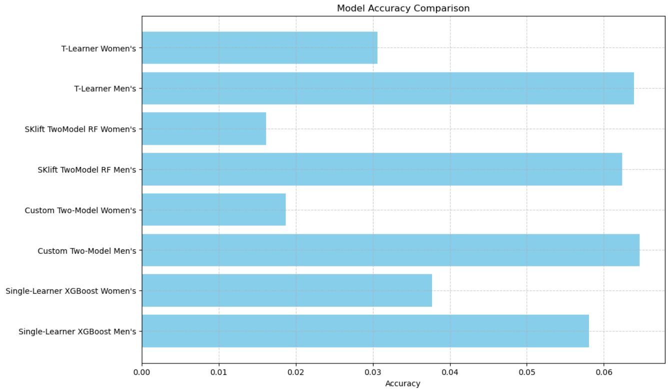


Fig. 8. Model Accuracy Comparison

- T-Learner models (Men's and Women's) show lower accuracy overall, reflecting their generally weaker performance in uplift as well.
- SKlift TwoModel RF (Men's) model presents the highest accuracy among all models, suggesting a strong predictive performance, which is consistent with its performance in the uplift curve.
- Custom Two-Model approaches (Men's and Women's) also demonstrate solid accuracy, which aligns with their strong performance in uplift.
- Single-Learner XGBoost (Men's and Women's) models exhibit a balanced performance, with accuracy in the middle range compared to other models.

E. Feature Importance Analysis

In uplift modeling, feature importance analysis is crucial for understanding which variables have the most significant impact on the model's predictions. Feature importance helps to identify the key drivers behind the uplift, allowing businesses to refine their targeting strategies [23]. In this section, we analyze the feature importance for two different models: the Single Learner XGBoost model and the T-Learner model, which includes separate models for the men's treated and control groups.

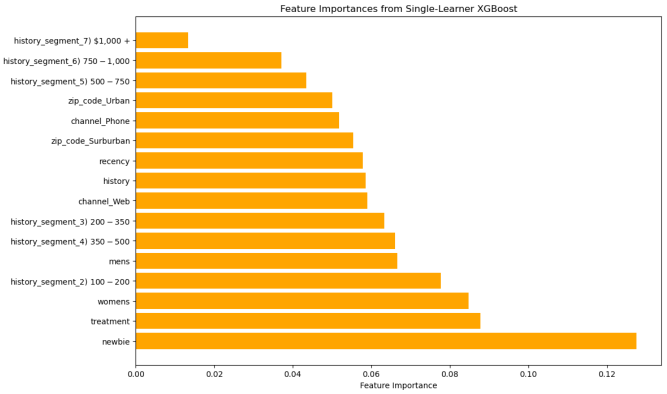


Fig. 9. Feature Importance for Single Learner (XGBoost)

1) Single Learner (XGBoost): The feature importance for the Single Learner XGBoost model is depicted in Fig. 9. The model's ability to handle complex interactions among features is evident from the distribution of feature importance. The newbie feature emerged as the most critical factor in predicting customer responses, followed closely by the treatment variable, which indicates whether the customer received the treatment. The features womens, mens, and various segments of purchase history (history) also showed significant influence, demonstrating their role in differentiating between different customer groups.

In a business context, understanding which features are most influential can help marketers focus on the most impactful factors when designing campaigns. For instance, the high importance of the newbie feature suggests that targeting new customers differently might yield better results.

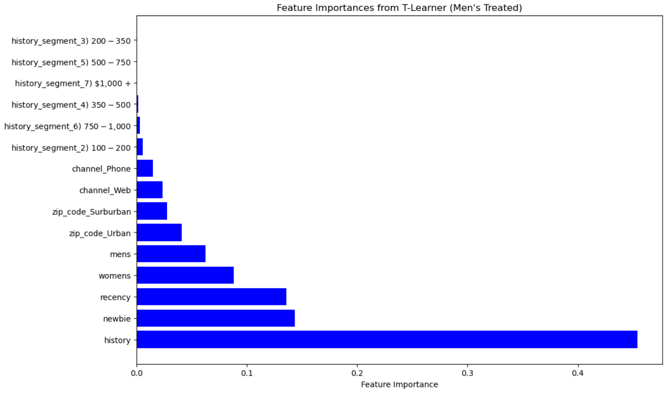


Fig. 10. Feature Importance T-Learner (Men's Treated)

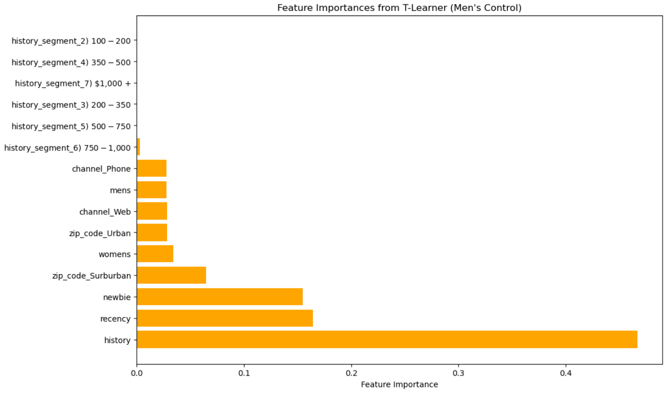


Fig. 11. Feature Importance T-Learner (Men's Control)

2) T-Learner (Men's Treated and Men's Control): The T-Learner approach involves separate models for the treatment and control groups, with feature importance analyzed individually for each. Fig. 11 shows the feature importance for the men's control group, while Fig. 10. represents the men's treated group. For the men's control group, the history feature stands out as the most influential, followed by 'recency' and 'newbie'. These features are essential for predicting the base-line behavior of customers who did not receive the treatment. On the other hand, in the men's treated group, the history feature remains dominant, but there is a noticeable shift in the importance of other features, such as 'recency' and womens, indicating how the treatment interacts differently with these variables.

The feature importance analysis across different uplift models provides critical insights into the drivers of customer behavior. The Single Learner XGBoost model highlights the overall significance of features like newbie and treatment, while the T-Learner models for men's treated and control groups reveal how these features influence behavior differently based on whether customers received the treatment or not.

F. Interpreting SHAP Values

In our experiment, SHAP (SHapley Additive exPlanations) was utilized to enhance the interpretability of our machine learning models. SHAP summary plots revealed the importance of individual features in driving model predictions [8], while dependency plots illustrated how changes in specific features influenced outcomes. These tools provided crucial insights into the model's decision-making process, ensuring transparency and explainability in the predictions made by the Single Learner and T-Learner models.

Reading SHAP Values:

- SHAP Summary Plot: This plot shows which features are most important for making predictions. If a feature has a lot of points far from zero on the x-axis, it means this feature has a big influence on the model's decisions. The color shows whether the feature value is high (red) or low (blue).
- SHAP Dependency Plot: This plot focuses on one feature and shows how changing its value affects the prediction. It also shows how this feature interacts with another feature. This is useful to see if two features combined have a bigger or smaller effect than expected.

These tools are powerful for interpreting complex models and understanding why they make certain predictions, making the machine learning models more transparent and trustworthy.

1) Single Learner (XGBoost): In the context of the Single Learner model, the SHAP (SHapley Additive exPlanations) summary plot provides a comprehensive overview of the impact each feature has on the model's output. In layman's terms, each point on the plot represents a single observation in the dataset, and the position of the point along the x-axis indicates how much the feature value influenced the model's prediction for that specific observation. The color gradient (from blue to red) indicates the feature value, with red points indicating high feature values and blue points indicating low feature values.

For instance, in the SHAP summary plot for the Single Learner model (Fig. 12.), we can see that features like "newbie," "recency," and "history" have the most substantial influence on the model's predictions. A positive SHAP value indicates that the feature pushes the prediction higher (towards 1, indicating the likelihood of a positive outcome), while a negative value pulls it lower (towards 0).

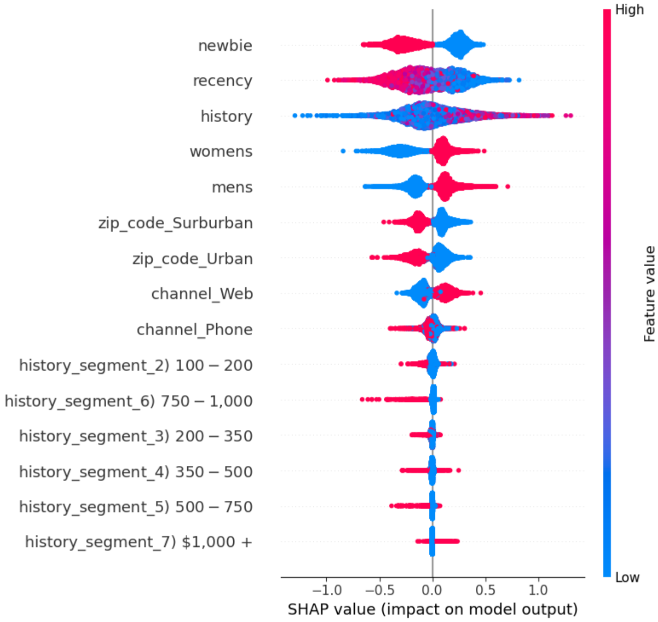


Fig. 12. SHAP Summary Plot - Single Learner

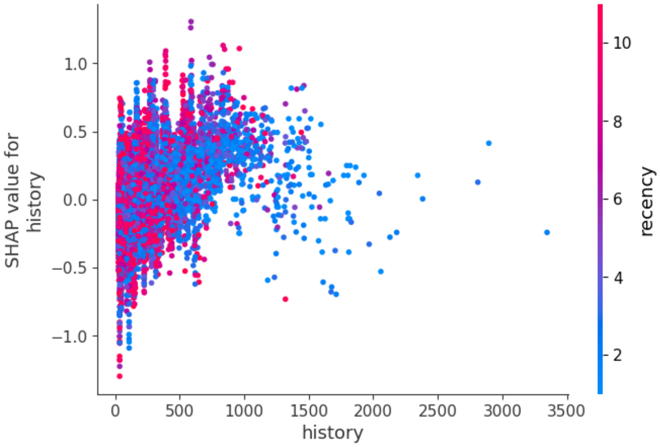


Fig. 13. Feature Dependency Plot - Single Learner

The SHAP dependency plot in Fig. 13., which we generated for the "history" feature, provides a deeper understanding of how this specific feature interacts with another feature (in this case, "recency"). The dependency plot reveals the relationship between the feature's value and its SHAP value, highlighting how changes in "history" influence the model's predictions while accounting for the interaction effect of "recency."

2) T-Learner (Gradient Boosting): For the T-Learner model, which involves separate models for the treatment and control groups, the SHAP summary plots allow us to compare feature importance between these two models. By examining these plots, we can identify how the importance of features like "history" and "recency" varies between the treated and

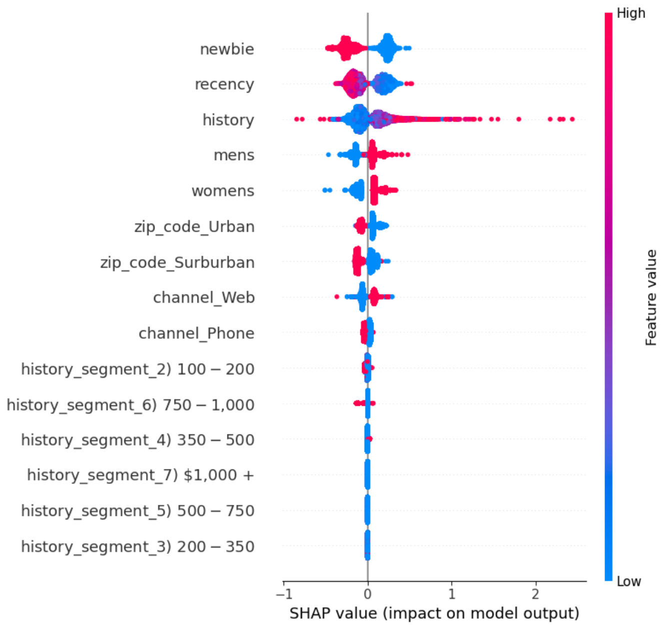


Fig. 14. SHAP Summary Plot - T Learner

control groups [17].

In the SHAP summary plot for the T-Learner's treated model (Fig. 14), the "history" feature again emerges as a key factor, but its influence might differ when compared to the control model. This comparison is crucial for understanding how the model interprets the effect of the treatment. The colors and spread of points give insights into how consistently a feature affects the predictions across different observations.

VI. Discussion

The main findings of this study reveal that all tested uplift models, including the Single Learner (XGBoost), Custom Two-Model, SKLift RF Two-Model, and T-Learner approaches, were effective in predicting the differential impact of targeted marketing campaigns on various customer segments. The Custom Two-Model approach exhibited the highest uplift scores, particularly for the men's campaign, suggesting its effectiveness in capturing nuanced differences between treated and control groups. However, the Single Learner approach, with its simplicity and efficiency, provided competitive results and demonstrated the ability to handle the complexity of interactions within the data. Our findings align with most studies done individually that ensemble methods like Random Forest and Gradient Boosting tend to perform well in uplift modeling [15], [16]. This is recommended that ensemble models are very effective tools for the prediction and generalization of high frequency big data [21]. However, the higher performance of the Custom Two-Model approach in this study underscores the importance of model selection based on the specific context and dataset characteristics.

One significant contribution of this study is the incorporation of explainability techniques, particularly SHAP values, to interpret the models' decisions. The SHAP summary and dependency plots provided valuable insights into feature importance and interactions, allowing for a more transparent understanding of how the models arrived at their predictions. The SHAP is the most powerful technique for interpreting predictive models, and the Shapley value represents the contribution weight of each variable to the prediction model. [19] This level of explainability is crucial for building trust in machine learning models, especially in business applications where decisions based on model outputs can have significant financial implications.

Despite the positive outcomes, there are limitations to this study. The reliance on the Hillstrom dataset, which is relatively small and specific to a particular industry, may limit the generalizability of the findings. Additionally, while the SHAP analysis provided insights into feature importance, it may not capture the full complexity of interactions within the models, particularly in more complex real-world datasets.

Future work could explore the application of these models to larger, more diverse datasets and consider the integration of other explainability techniques [4], such as LIME or counterfactual explanations, to further enhance model transparency. These interpretable models can consequently be used to explain the prediction that is made by the actual, black-box model. [20]. Moreover, the extension of these models to other domains, such as healthcare or finance, could provide valuable insights into their applicability and effectiveness in different contexts.

The study highlights the importance of model selection and explainability in uplift modeling, offering a robust approach for targeted marketing and other personalized interventions. The findings have practical implications for businesses seeking to optimize their marketing strategies, as well as for the broader field of machine learning, where transparency and trust are becoming increasingly important.

VII. Conclusion

In this study, we explored and compared various uplift modeling techniques, including Single Learner (XGBoost), Custom Two-Model, SKLift RF Two-Model, and T-Learner, within the context of targeted marketing campaigns. Our results demonstrated that the Custom Two-Model approach yielded the highest uplift scores, particularly for the men's campaign, emphasizing its effectiveness in differentiating treatment effects. However, the Single Learner model proved to be a competitive and efficient alternative.

The integration of explainability through SHAP values provided transparency, revealing key feature influences and interactions, thus enhancing the trustworthiness of the models. While the study's findings offer valuable insights, the limitations related to dataset specificity and model generalizability suggest areas for future research. Overall, this work contributes to the field by highlighting the importance of model selection and explainability in uplift modeling, with practical implications for optimizing marketing strategies.

VIII. Declarations

A. Declaration of Originality

I am aware of and understand the University of Exeter's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices.

B. Declaration of Ethical Concerns

This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also no security or safety critical activities have been carried out.

References

[1] Devriendt, F., Berrevoets, J. and Verbeke, W., 2021. Why you should stop predicting customer churn and start using uplift models. Information Sciences, 548, pp.497-515. Available at: https://doi.org/10.1016/j.ins.2019.12.075.

[2] Zhang, W., Li, J. and Liu, L., 2021. A unified survey of treatment effect heterogeneity modelling and uplift modelling. ACM Computing Surveys, 54(8), Article 162, pp.1-36. Available at: https://doi.org/10.1145/3466818.

[3] Radcliffe, N.J. and Surry, P.D., 2012. Real-world uplift modelling with significance-based uplift trees. In Proceedings (no specific conference mentioned). Available at: https://api.semanticscholar.org/CorpusID:17521088.

[4] Burkart, N. and Huber, M.F., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, pp.245-317. Available at: https://doi.org/10.1613/jair.1.12228.

[5] Sołtys, M., Jaroszewicz, S. and Rzepakowski, P., 2015. Ensemble methods for uplift modeling. Data Mining and Knowledge Discovery, 29, pp.1531-1559. Available at: https://doi.org/10.1007/s10618-014-0383-9.

[6] Devriendt, F., Moldovan, D. and Verbeke, W., 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data, 6(1), pp.13-41.

[7] Ganjisaffar, Y., Caruana, R. and Lopes, C., 2011. Bagging gradient-boosted trees for high precision, low variance ranking models. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 85-94). Available at: https://doi.org/10.1145/2009916.2009932.

[8] Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2021. Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), p.18. Available at: https://doi.org/10.3390/e23010018.

[9] Gubela, R.M. and Lessmann, S., 2021. Uplift modeling with value-driven evaluation metrics. Decision Support Systems, 150, p.113648. Available at: https://doi.org/10.1016/j.dss.2021.113648.

[10] Zaniewicz, Ł. and Jaroszewicz, S., 2013. Support vector machines for uplift modeling. In 2013 IEEE 13th International Conference on Data Mining Workshops (pp. 131-138). Dallas, TX, USA: IEEE. Available at: https://doi.org/10.1109/ICDMW.2013.23.

[11] Gutierrez, P. and Gérardy, J., 2017. Causal inference and uplift modelling: A review of the literature. In Proceedings of The 3rd International Conference on Predictive Applications and APIs, Proceedings of Machine Learning Research, 67, pp.1-13. Available at: https://proceedings.mlr.press/v67/gutierrez17a.html.

[12] Künzel, S.R., Sekhon, J.S., Bickel, P.J. and Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences of the United States of America, 116(10), pp.4156-4165. Available at: https://doi.org/10.1073/pnas.1804597116.

[13] Guelman, L., Guillén, M. and Pérez-Marín, A.M., 2015. Uplift random forests. Cybernetics and Systems, 46(3-4), pp.230-248.

[14] Sun, C., Li, Q., Wang, G., Xu, S. and Liu, Y., 2023. KDSM: An uplift modeling framework based on knowledge distillation and sample matching. arXiv preprint arXiv:2303.02980.

[15] Nielsen, D., 2016. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? (Master's thesis, NTNU).

[16] de Jongh, A., ten Broeke, E., & Meijer, S. (2010). Two Method Approach: A Case Conceptualization Model in the Context of EMDR. Journal of EMDR Practice and Research, 4(1), 12–21. https://doi.org/10.1891/1933-3196.4.1.12

[17] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, pp.1937-1967. Available at: https://doi.org/10.1007/s10462-020-09896-5.

[18] Zhao, Z. and Harinen, T., 2020. Uplift modeling for multiple treatments with cost optimization. arXiv. Available at: https://arxiv.org/pdf/1908.05372.

[19] Nohara, Y., Matsumoto, K., Soejima, H. and Nakashima, N., 2019. Explanation of machine learning models using improved Shapley additive explanation. In Proceedings of the 2019 ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 546-546). Available at: https://doi.org/10.1145/3307339.3343255.

[20] Brughmans, D., Melis, L. and Martens, D., 2024. Disagreement amongst counterfactual explanations: how transparency can be misleading. TOP. Available at: https://doi.org/10.1007/s11750-024-00670-2.

[21] Shrivastav, L. and Kumar, R., 2022. An ensemble of random forest, gradient boosting machine, and deep learning methods for stock price prediction. Journal of Information Technology Research, 15, pp.1-19. Available at: https://doi.org/10.4018/JITR.2022010102.

[22] Karlsson, H., 2019. Uplift modeling: Identifying optimal treatment group allocation and whom to contact to maximize return on investment [dissertation]. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-157962

[23] Bamidele, T. and Mgbaja, U., 2024. Enhancing targeted marketing strategies: Interpretable uplift modeling to identify key client segments. [Journal Name if available]. Available at: https://doi.org/10.21203/rs.3.rs-4006839/v1.