Effectiveness of Uplift Modeling When Multiple Treatments are Tested in Fashion E-Commerce Campaigns

MSc Research Project
Data Analytics

Prajwal Keshav Kongi
Student ID: 22205314

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Prajwal Keshav Kongi
Student ID:	22205314
Programme:	Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	12/08/2024
Project Title:	Effectiveness of Uplift Modeling When Multiple Treatments are Tested in Fashion E-Commerce Campaigns
Word Count:	6900
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Effectiveness of Uplift Modeling When Multiple Treatments are Tested in Fashion E-Commerce Campaigns

Prajwal Keshav Kongi
22205314

Abstract

The fashion e-commerce sector relied on generalized marketing methods like sending promotional messages to every customer, leading to customer fatigue and ineffective campaigns. This study aims to apply and evaluate multiple uplift modeling techniques to target potential customers in the e-commerce domain. The focus is on identifying 'Persuadable' customers, those who are likely to respond positively to marketing messages sent through SMS, Email, or Push Notifications. The study is done on a Russian medium-sized online fashion retailer and represents customer's interaction with campaigns from 2021-2023. The message clicked by the customer is considered the target action in this study. Four uplift models are implemented, the Two-Model Approach with CatBoost and LightGBM classifiers, Class Transformation, and T-Learner with LightGBM. The metrics used for the evaluation are Uplift Score at 30%, AUUC, and AUQC. Class Transformation achieved the best Uplift Score at 30% with 7%, indicating moderate success in targeting potential customers. Generally, the overall performance across the models was mixed, with low AUUC and AUQC scores. Signifying challenges in generalizing the models across the data. The research suggests uplift modeling be applied to enhance targeted marketing in fashion e-commerce. However, the study showed several key limitations such as computational complexity, and difficulties with capturing complex customer behaviors.

1 Introduction

This project aims to target potential customers for fashion e-commerce domain by using uplift modeling techniques. E-commerce companies have depended on generalized marketing strategies in the past. Targeting or showing advertisements to all the customers hoping to find potential customers. But now, companies need to identify potential customers and convert them effectively i.e, to make them purchase their products or achieve a targeted action from the customers. Modern e-commerce platforms use Push notifications, Emails, and SMS as means to communicate with their customers. The problem with traditional marketing is that with abundant promotional messages, it leads to message fatigue for the customers. The challenge is not only in optimizing the content of the messages but to target and send the messages to only the potential customers to ensure cost efficiency.

The motivation of the research and details about the uplift modeling techniques applied on the fashion e-commerce domain is described in Section 1.1.

1.1 Motivation and Uplift Modeling

The motivation of this project is to apply uplift modeling techniques in the fashion e-commerce industry to target potential customers when multiple treatments are tested. Multiple treatments in this study are sending promotional, discount and other marketing related ads through multiple channels (Push notification, Emails, and SMS). There is limited research of uplift modeling in e-commerce domain. This research explores uplift modeling as an answer to the limitations of traditional predictive analytics. Traditional machine learning models can predict whether a customer is going to buy something, but they fall short of determining the real impact of a marketing campaign intervention. Whereas, uplift modeling is concerned with the causal inference of the effect of treatment (such as a marketing message) on individual customer outcomes.

The uplift model categorizes customers into four groups based on their response to marketing interventions: Do-Not-Disturbs, Lost Causes, Persuadable, Sure Things. The focus of this project is primarily on identifying and targeting the "Persuadable" group, customers who are likely to make a purchase or do desired action only if they receive marketing messages. The dataset used in this study is sourced from a medium-sized Russian online fashion e-commerce company from 2021-2023. It contains data on marketing messages sent to customers via multi-channel i.e, through Push notifications, SMS and Email. The key features in the data are messages sent date, message content, message opened flag, message clicked flag, and item purchased flag. The average purchase order value is $60.

In this study, the target action is whether the message was clicked or not. This is selected as the target action and not whether the customer purchased the product because the conversion rate is below 2%, which is considered good in fashion e-commerce industry. But in this study, we are checking for whether the customer clicked the message or not. The click rate is around 20%. Four uplift modeling approaches were implemented, compared and evaluated. Two-Model Approach with CatBoost classifier, and with LightGBM classifier. Class Transformation and Meta Learner uplift models are implemented. The uplift score at 30% for all the model are around 0.15%, except the Class Transformation model, which achieved a higher uplift score of 7% among all the other models (i.e, top 30% customers clicked the message). The other evaluation metrics include AUUC and AUQC where all the models resulted marginal scores, suggesting minimal overall impact. Despite incorporating regularization and other optimization techniques, all the models struggled to generalize across the entire dataset, providing low results. This emphasizes the complexity of understanding customer behavior and the need for further refinement in the uplift modeling approaches.

Currently, there is limited research on uplift modeling when multiple treatments are to be tested in the fashion e-commerce domain. Therefore the following research question is presented in this project.

Research Question: "How effective are uplift modeling techniques in predicting customer response to marketing campaigns between treatment and control groups within the fashion e-commerce domain in Russia, particularly when multiple treatments such as discounts and personalized recommendations are tested simultaneously across channels like Email, SMS and Push notifications?"

To answer the research question, Section 1.2 is implemented for this project which describes the contribution and objectives of the research.

1.2 Objectives and Contributions

Table 1 describes the contribution and objectives of the research.

ID	Description
Obj. 1	A critical analysis of the literature on Uplift modeling in the fashion e-commerce domain when multiple test scenarios are present.
Obj. 2	Implementation, and evaluation of the different uplift modeling techniques. The proposed models are evaluated and results are provided in the Obj. 3.
Obj. 3	Comparison of the developed uplift models. The results of the models provide a better-suited model in the fashion e-commerce domain when multiple test scenarios are present.

Table 1: Objectives of the Research

In conclusion, the research explores and compares various uplift modeling techniques and determines the most suitable model for the fashion e-commerce domain. To provide practical recommendations for businesses to enhance the effectiveness of their marketing strategies using uplift modeling. Section 2 explores the related works on uplift modeling within the e-commerce domain, specifically focusing on techniques where multiple treatments are evaluated.

2 A Review of Uplift Modeling in E-Commerce Multi-Channel Campaigns

2.1 Uplift Modeling and Evaluation Methods

A dual uplift model and an alternative XGBoost model combined enhanced the prediction accuracy in the study conducted by Kodikara and Shahtahmassebi (2023). In this case, the models are quite useful and have good performance in segmenting customers into different categories, such as Persuadables and Lost Causes, considering the targeted marketing. Similarly, in the work by Singh et al. (2023), while the main dataset comprised RCTs that can serve as a good representation of real-world marketing campaigns, significant results were reported when optimizing customer targeting using uplift modeling. However, it was noted that the techniques generally suffered from overfitting challenges and consumed substantial computational resources, particularly when hypercomputing techniques were integrated. Another study by Zaniewicz and Jaroszewicz (2013) is on the adaptation of support vector machines for uplift modeling. The aim of the study was to predict the differential impact of treatment across different groups. By reformulating the SVM optimization problem, the study addresses the challenge of identifying positive, neutral, or negative responses to a given action. For example, researchers made the evaluation of their model using publicly available datasets, such as e-mail campaigns and clinical trial data. The Uplift SVM has already shown good, sometimes even superior results in comparison to common benchmark methods, like the two-model approach. This research is particularly relevant to our study, as it showcases the importance of using uplift modeling to optimize marketing strategies in e-commerce. This provides a foundation for further exploration of multi-treatment uplift models in the fashion industry.

To support coupon advertising that is more cost-effective on a C-to-C e-commerce platform, uplift modeling was used in the study by Shimizu et al. (2019). The authors ran various models, such as SGD, Random Forest, and XGBoost, using 26 days of customer log data to reach the best model approach. Indeed, results showed better performance from the XGBoost model, with a 39% reduction in marketing costs and near-negligible losses in the number of customers acquired. Another study is by Wang, Xu, Feng, Ignatius, Yin and Xiao (2024), which proposed a new method, PSM-NDML. By combining the propensity score matching with double machine learning, an estimation of causal effects is made with respect to the delivery appointment's impact on success. The method is applied in a real-world dataset of a Chinese logistics company and competes well with the best uplift methods. It scores highest in both the gain and Qini coefficient metrics. These two studies remain relevant to our study, they apply uplift modeling for real-world e-commerce marketing scenarios, so they hold the potential for realizing significant cost-saving. In contrast, the implementation of these advanced models, such as the XGBoost, over larger data sets, is quite complex. Future research could explore how scalable these methods are to e-commerce on a much broader scale.

Uplift modeling techniques have recently been applied to B2B customer churn prediction following a segmentation-based modeling methodology De Caigny et al. (2021). The research considered a real-world dataset of 6,432 customers from a software provider in Europe and built models based on uplift decision trees and logistic regression models. In general, the main results showed that the proposed uplift logit leaf model was outperforming others in terms of predictive performance. However, a drawback is that it works with a single dataset and that there is the danger of overfitting in complex models.

It is essential to evaluate uplift models. Variance reduction techniques may help in evaluating the uplift model, especially in RCTs. Bokelmann and Lessmann (2024) achieved this by adjusting outcome predictions and significantly reducing the variance in metrics like the Qini curve, enhancing the precision of model assessments. Also enhancing metrics like AUUC and AUQC scores. Still, these methods introduce complexities that require their careful application. The development of a framework for the generation of Surrogate Ground Truth (SGT) that improves binary fairness evaluation in uplift modeling is explained in the research by Michalský and Kadıoğlu (2021). This procedure, when ground truth labels are not at disposal, generates surrogate labels that approximate the counterfactual outcomes. The efficiency of SGT can be proved by the study of real marketing campaigns and is especially effective in bettering fairness evaluations. However, it can introduce bias from the initial uplift model, and then the accuracy of the surrogate labels may vary with the quality of the initial predictions.

The researchers propose new evaluation metrics that consider both the individual treatment effect (ITE) and the expected customer value to optimize the targeting process in this study Gubela and Lessmann (2021). This is validated on different real-world datasets derived from a number of e-commerce campaigns. The study indicates that the profit metrics are significantly improved over traditional ITE-based methods. However, one limitation is the assumption of linear relationships in their model, which might oversimplify real-world complexities. While the study in the paper introduces novel metrics to enhance profit, our study applies and compares multiple uplift models directly. The paper's methodology could inspire future improvements in our approach, particularly in integrating profit-based evaluation metrics. Similarly, the study by Gubela et al. (2020) explores the application of revenue uplift modeling. By focusing on the causal effects of marketing treatments on customer revenue rather than mere conversion rates, the findings demonstrate significant improvements in targeting and revenue optimization.

2.2 A Review on Uplift Modeling when Multiple Treatments are Involved

The effectiveness of uplift modeling with a Siamese neural network is researched by Peralta et al. (2023). This study is done on a retail company in Chile which points toward better performance of the model in predicting customer responses. Importantly, in this research study, the evaluated e-commerce platforms handle multi-channel communications similar to our current study such as email, SMS, and push notifications. The researchers show that the Siamese neural network model significantly outperforms the classical uplift model such as the Two-Model approach in the Qini curve metric, with an area under the curve 2.5 times larger. The study by Ke et al. (2021) addresses the challenge of exposure bias, able to compromise the true estimation of the causal effect of ad exposure. The authors, therefore, set out to adjust this bias through an Explicit Uplift Effect Network (EUEN) and its extension, the Explicit Exposure Uplift Effect Network (EEUEN) for achieving a more accurate uplift model. The study shows significant improvements in performance metrics such as AUUC curve and the Qini coefficient. In comparison, we have employed similar uplift models like Two-Model and Class Transformation, achieving lower uplift scores and minimal AUUC and AUQC values. This highlights the potential benefit of sophisticated neural networks in fashion e-commerce marketing campaigns. It becomes evident that the result obtained is largely influenced by the model chosen and specific context (such as online advertising). The models we tested in the research, with metrics such as uplift at 30%, performed poorly compared to the advanced techniques used in this study.

The M3TN is a Multi-gate Mixture-of-Experts based Multi-valued Treatment Network, developed in the study by Sun and Chen (2024). For treating inefficiencies in uplift modeling for multiple treatments. This model combines a feature representation module with reparameterization, resulting in enhanced efficiency and effectiveness. The M3TN always outperformed the existing models, especially in the case of multiple treatments. The strengths of the M3TN model are primarily due to its architecture, which minimizes cumulative errors and improves overall model efficiency. However, complexity will be a limiting factor for wider use in environments where resources are lower. Additionally, Devriendt et al. (2022) explores the application of Learning to Rank (L2R) techniques in uplift models. The researchers concentrated on the optimization of ranking of instances according to estimated ITE values, not ITE values themselves. They introduced a new metric, the Promoted Cumulative Gain (PCG), and applied the LambdaMART L2R method in order to improve the performance of uplift modeling. The results showed that methods based on L2R, mainly when adapted for specific measures of ranking, such as the PCG, give better results than traditional uplift models. However, a key limitation is that these improvements may not consistently generalize well to new data.

A Generalized Causal Tree (GCT) algorithm for uplift modeling is developed in the study by Nandy et al. (2023), which can handle multiple discrete or continuous treatments. This approach tries to improve the traditional tree-based methods by overcoming the limitations of the earlier algorithms, which have been more concentrated on binary treatment and not multiple treatments like our case. The GCT algorithm is most useful in applications where optimum treatment allocation for diverse populations is required. However, the implementation cost and complexity of handling treatments can be challenging.

2.3 A Review on Advanced Uplift Modeling Techniques and Novel Approaches

E-commerce marketing has transformed from general approaches to more personalized strategies, by using models such as uplift modeling Sato et al. (2019). This modeling technique is essential for optimizing multi-channel campaigns in the fashion e-commerce industry, where customer engagement can be significantly influenced by targeted communications through SMS, push notifications, and emails. The study by Wang et al. (2022) introduces the Multihead Causal Distilling Weighting (MCDW) method, a novel approach to uplift modeling that emphasizes feature selection based on causal inference to ensure the interpretability of the model and eliminate overfitting risks. These methods, compared to the traditional approaches, reduce most of the complexity and cost associated with feature control, as it mainly focuses on accuracy as its major objective.

The RUAD framework incorporates a feature selection module into the joint multi-label modeling to make uplift models more stable and accurate across datasets. In the paper by Sun et al. (2023), a novel approach to enhance uplift modeling by increasing its robustness is proposed through a framework, for Robustness-enhanced Uplift Modeling with Adversarial Feature Desensitization (RUAD). This served as a quantitative demonstration that is effective in maintaining model performance. This method addresses the sensitivity of uplift models to key features, which is crucial to the performance of uplift models. The most important part of this paper is to increase the robustness of the uplift models. Therefore, when applying these models to dynamic environments, e.g., online marketing, the results are strongly enhanced. Also the study Wang, Ye, Wang, Zhou, Zhang, Zhang, Jiang and Zhai (2024) introduces a Graph Neural Network framework with causal knowledge to uplift model enhancement. This work uses both synthetic and real-world data sets. The model tested against the CRITEO dataset achieved an AUUC score of 0.8807, which is significant when compared to classical models like linear regression or XGBoost. However, the complexity of the GNN framework might be a challenge during real-world implementations. This helped to deal with the confounding issues across mixed treatments and estimate the CATE for each treatment.

A novel uplift-modeling technique called weighted doubly robust learning (WDRL) was developed by Zhan et al. (2024). This method combines the Shapley-value treatment attribution with doubly robust estimation to control the confounding effects. Leading to superior customer targeting and treatment attribution relative to traditional uplift models. In the study, the AUUC score was 0.590, and the Qini-coefficient at 0.080, far more significant than existing models. Comparing this with our research, the best model achieves an uplift score at 30% of only 0.0709. The discrepancy in results between this and our study probably lies in the complexity of real-world datasets and overfitting problems that may be available in the base learners used in our research.

A novel method called propensity score oversampling and matching (ProSOM) for confounding in uplift modeling is experimented by Vairetti et al. (2024). The experiment was conducted using datasets from various domains, including finance and retail. To show the effectiveness of ProSOM in uplift modeling for improvement in predictive performance. For most datasets, results showed that ProSOM beat traditional methods like PSM and other resampling techniques concerning Qini coefficients and AUUC. The approach can also be biased towards the minority class and can be very computationally heavy, so more tuning is needed.

The Unbiased Contextual Treatment Selection (UCTS) algorithm was proposed by Zhao et al. (2017). With the main purpose of increasing both uplift model accuracy and reliability. An important aspect of the UCTS algorithm is its ability to separate the feature space partition from leaf response estimation, addressing biases that can compromise the effectiveness of other algorithms. The major contribution of this study is to showcase that the proposed UCTS performs better on synthetic and real-world data. However, the limitations include the sensitivity of hyperparameter tuning which results in some drawbacks, as generalization of success with different datasets is not guaranteed. We encountered similar challenges, as the performance of the models varied depending on the dataset's characteristics.

2.4 Conclusion

In conclusion, there has been advances in uplift modeling but limited research in the fashion e-commerce domain especially when multiple treatments are to be tested via communication channels (Email, SMS and Push notifications). Uplift models hold great potential for the improvement of targeted marketing strategies in many areas. In turn, they bring along issues of overfitting, computational complexity, and scalability. Section 3 aims to discuss the methodology of uplift modeling on retail fashion e-commerce in Russia where multiple treatments are being tested.

3 Methodology

Cross-Industry Standard Process for Data Mining (CRISP-DM) is one of the powerful and frequently used methodologies that allows a structured approach to a data mining project. In this research, we are going to implement the CRISP-DM framework which will guide the general process of data mining. CRISP-DM follows six major phases and each phase will be elaborated as follows.

3.1 Business Understanding

This step involves understanding the business aspect of the research. The motivation and goal of the project are defined. In a competitive industry such as fashion e-commerce, companies have been using generalized marketing strategies, broadcasting promotional messages to the masses with the hope of targeting potential customers. Yet, such an approach often results in message fatigue for consumers, thereby reducing the effectiveness of marketing campaigns and increasing costs. The general goal of the project is to improve customer targeting with uplift modeling methods that identify and convert potential customers who are most likely to respond positively to marketing interventions. By targeting potential targets to show marketing ads, companies can reduce unnecessary costs that go into customers who do not respond, increase the rate of engagement, and consequently heighten the campaign effectiveness of marketing.

3.2 Data Understanding

The data understanding phase involves exploring the dataset. The dataset contains messages obtained from a medium-sized Russian online fashion e-commerce company, spanning two years from 2021 to 2023. The analysis of the data is performed in Jupyter Notebook. The dataset has over 10 Million records. For this research, we are working on 1 Million records because of limitations in computational resources. The average purchase order value is $60. The dataset is divided into 2 files. Messages data and Marketing Campaigns data.

The dataset contains a variety of features that are essential for analyzing customer behavior. Key features include:

- Campaign Information: Includes the campaign_id, which uniquely identifies each marketing campaign.
- Customer Interaction Data: Every message contains detailed statistics about delivery, open, click, purchase, and all negative events like unsubscribe, spam complaint, and bounce features with flags and date time.
- Treatment Indicator: The treatment column indicates whether a customer was part of the treatment or control group, which is vital for uplift modeling.
- Message Characteristics: Features like subject length, subject with personalization, and subject with emoji provide details about the content of the messages sent to customers.
- Channels: Multichannel means they send messages via different channels: email, web push, mobile push, and SMS.

Initial exploratory data analysis revealed that the click rate is around 20%, and the conversion rate is below 1%, which aligns with industry standards in fashion e-commerce. We have not considered the purchase as our desired action for this research. This research focuses on predicting the likelihood of a customer clicking on a marketing message as the target action.

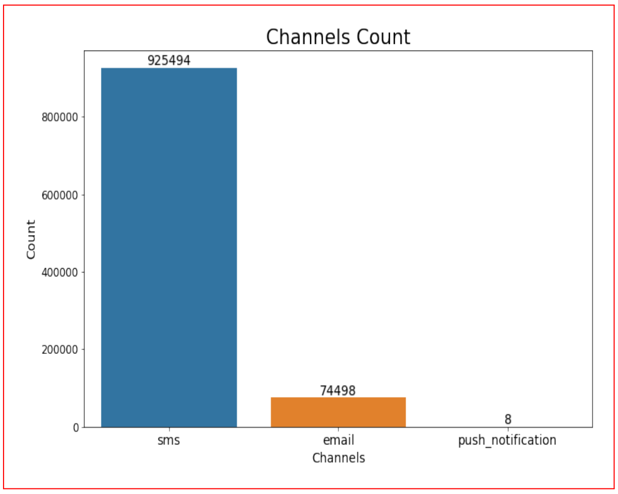

Figure 1: Distribution of Channels

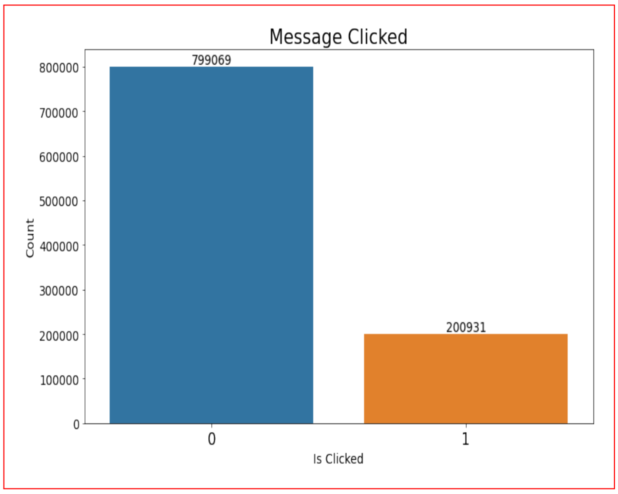

Figure 2: Distribution of Messages Clicked

From Figure 1, we can see the distribution of the channels feature indicating a strong preference for SMS, with 925,494 instances, followed by Email with 74,498, and minimal use of Push notifications. From Figure 2, the message clicked feature (target action) shows that 19.97% of the total messages sent to the customers resulted in a click, highlighting the effectiveness of the campaigns in generating user engagement.

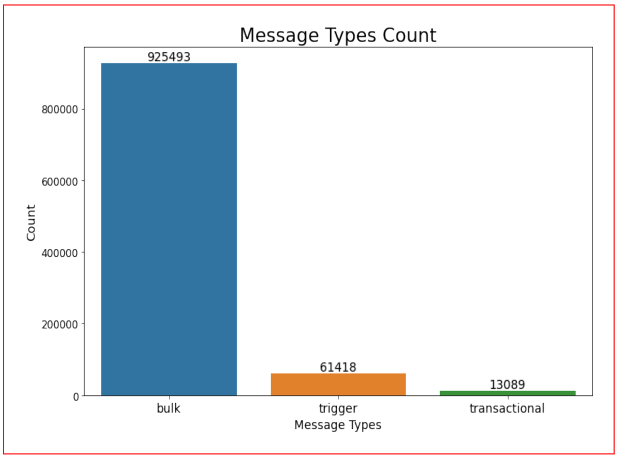

Figure 3: Distribution of Message Types

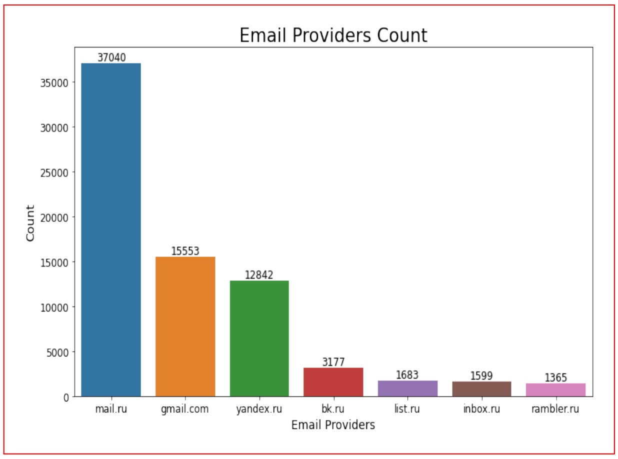

Figure 4: Distribution of Email Providers

From Figure 3, the distribution of the Message Type feature reveals that the bulk of the messages were of the "bulk" type, totaling 925,493. The difference between bulk and two other types is that Bulk campaign messages can be sent multiple times. Triggers and transactions are sent only once. That's why the dataset has a lot of bulk campaigns and a small number of other types. In Figure 4 of Email Providers distribution, "mail.ru" is dominating with 37,000 instances followed by "gmail.com". This indicates that these are the primary email domains used by customers in Russia. Mail.ru is the largest webmail provider in Russia with 100 million active email accounts.

3.3 Data Preparation

In the Data Preparation phase of CRISP-DM, the primary goal is to transform and clean the dataset to ensure it is suitable for modeling. The following methods were employed to preprocess the data, making it ready for the uplift modeling techniques applied in this project.

3.3.1 Fixing datatypes

The first transformation in the data preparation process was fixing datatypes of the features. Features like "campaign_id" needed to be converted from a string to an integer format. Features like "is_opened", and "is_clicked" were originally stored as boolean variables ('t', 'f'), they have been transformed to boolean integers (0 or 1). Lastly, the date-related columns were converted to datetime format for proper time-based calculations, particularly creating recency features that are key in customer behavior analysis and response timing for marketing campaigns.

3.3.2 Handling Missing values

For the missing values in Campaigns data. Categorical features were filled with their respective modes. Meanwhile, for the continuous features, such as "subject_length" and "total_count" of the subject line. These features's missing values were imputed with their respective means. Mean imputation is used in order to retain the statistical properties of the data, making all records available for modeling.

3.3.3 Aggregating Email Providers

There were initially 78 unique values in the 'email_provider' column. The top 7 email providers were kept, and the rest were converted to 'others'. This is because the other email provider's count was low, only the top 7 email providers' count was higher. This method is also helpful in dimensionality reduction when we apply one-hot encoding to the categorial features.

Further data processing like Feature Engineering steps is done in the Implementation Section 5.2.

3.4 Modeling

In this phase, uplift models are designed, trained, and tuned with various parameters to refine the models. In this study, four uplift models are implemented. Two-model approach with CatBoostClassifier, LGBM Classifier, Class Transformation with CatBoost, and T-Learner with LightGBM Classifier. The primary goal of this phase is to develop a model that can accurately target potential customers in the e-commerce domain. Further explanations of these models are discussed in detail in Implementation Section 5.3.

3.5 Evaluation

The evaluation metrics to evaluate uplift models are Uplift Score at 30%, Area Under the Uplift Curve (AUUC), and Area Under the Qini Curve (AUQC).

Uplift Score at 30%: It is a metric used to assess the uplift model's performance within the top 30% of customers, arranged by predicted uplift. The top 30% are then identified as the ones most likely to show the positive effect of treatment according to model prediction. This can be calculated as the response rate difference between the treatment and control groups in this top 30% segment.

Area Under the Uplift Curve: The value is computed as the area under a random uplift curve normalized by the area under the ideal uplift curve. The cumulative uplift in response rates of the population, ranked by predicted uplift, is given by the uplift curve.

Area Under the Qini Curve: It is similar to the uplift curve but it accounts actual distribution of treatments within the population. It plots the cumulative net benefit from targeting people in the treatment group over the control group, calculated as cumulative differences in positive outcomes within the group, normalized by the total population size.

The evaluation of implemented uplift models is discussed in Section 6.

3.6 Deployment

The final phase of CRISP-DM is deployment, which focuses on the final stage of the data mining project. Here, the developed model is implemented in a production environment. However, the scope of deployment is not included in this study. The primary focus was on developing and evaluating uplift modeling techniques to enhance customer targeting within the fashion e-commerce domain. The deployment stage of this study is for future work.

4 Design Specification

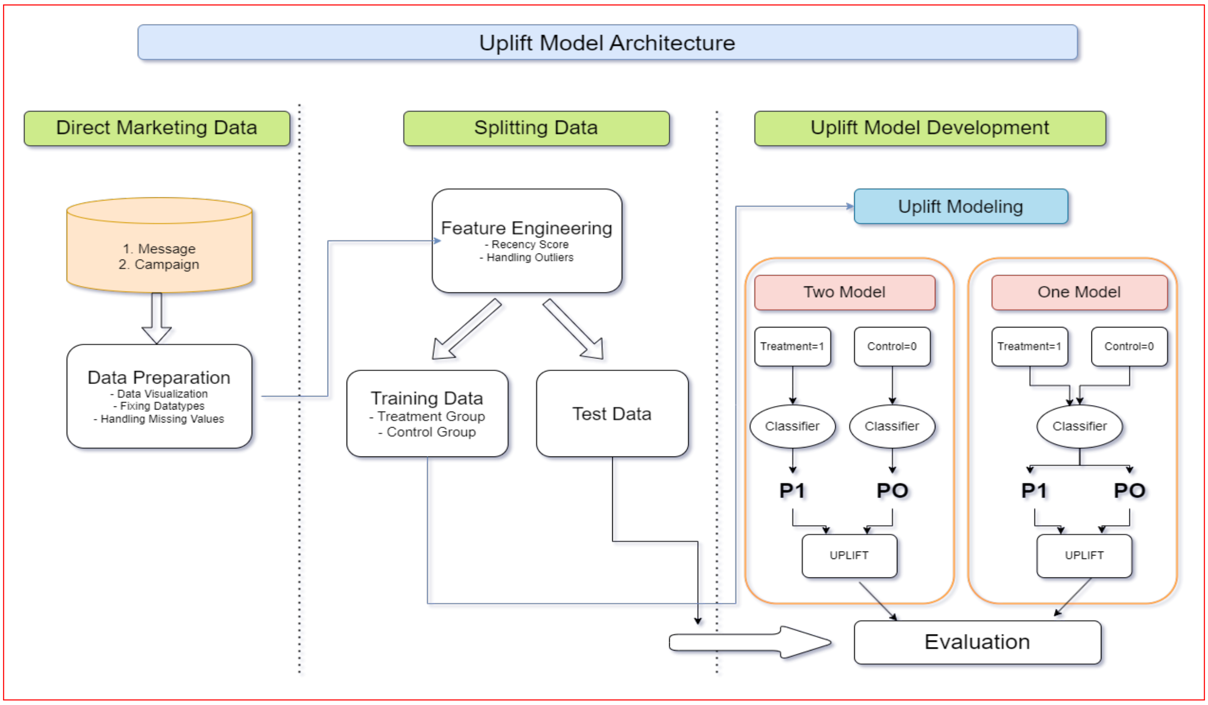

Figure 5: Architecture of Uplift Modeling

Figure 5 represents the architecture for implementing uplift modeling, particularly focusing on a fashion e-commerce domain to target potential customers. The architecture is divided into several key stages, each playing a crucial role in the uplift modeling process. The architecture starts with sourcing data of messages and campaign data. The Next step consists of Data Preparation such as Fixing datatypes and Handling Missing values as seen in Sub-section 3.3. Further Feature Engineering methods are implemented in Sub-section 5.2. The data is split into Training (Treatment and Control) and Test data. The training data is then trained on two types of uplift models. Finally, the performance of the trained model is evaluated on the Test data and the models are retrained with hyperparameter tuning until optimum results are achieved. Section 5 outlines the implementation of uplift modeling in detail.

5 Implementation

This section involves implementation of the proposed uplift modeling techniques along with feature engineering steps, and detailing the tools and technologies used.

5.1 Tools and Technologies

The project is implemented by utilizing Jupyter Notebook. Python was chosen for its extensive packages and libraries. Pandas and NumPy are the libraries used for data manipulation and preprocessing. These libraries are efficient in handling and transforming large datasets. Matplotlib and Seaborn were used for data visualization. The 'scikit-learn' library was used in data splitting, imputation, and construction of a pipeline since this is useful in ensuring the workflows run smoothly from data preprocessing steps up to model evaluation. The key library for implementing uplift modeling is 'scikit-uplift', which contains Two-Models and Class Transformation models. Additionally, 'CatBoost' and 'LightGBM' packages were chosen for modeling.

5.2 Feature Engineering

Feature engineering is the process of transforming raw data or creating new data from raw data to improve the performance of the model. In this study, the following 3 feature engineering steps are applied to the dataset.

5.2.1 Detecting and Treating Outliers

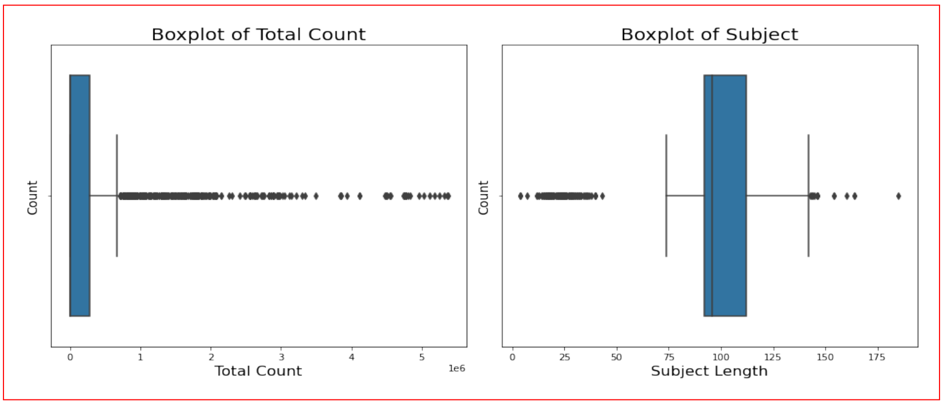

Figure 6: Boxplot: To Detect Outliers

Figure 6 shows the boxplot to detect outliers of the features. "Total Count" boxplot shows that the distribution of data is right-skewed. There are several significant values on the right side of the whisker, which indicates extreme values or anomalies. The box plot of "Subject Length" has symmetrical distribution of the data. However, we could see outliers on both sides.

In Figure 7, we can see the boxplot after treating outliers. For "Total Count", the Log Transformation method is used to normalize the data. This is because it reduces skewness as there are a large number of outliers towards the right side. For the "Subject Length", the Cap Outliers method is used based on the 5th and 95th percentiles. This mitigates the impact of these extreme values on both sides.

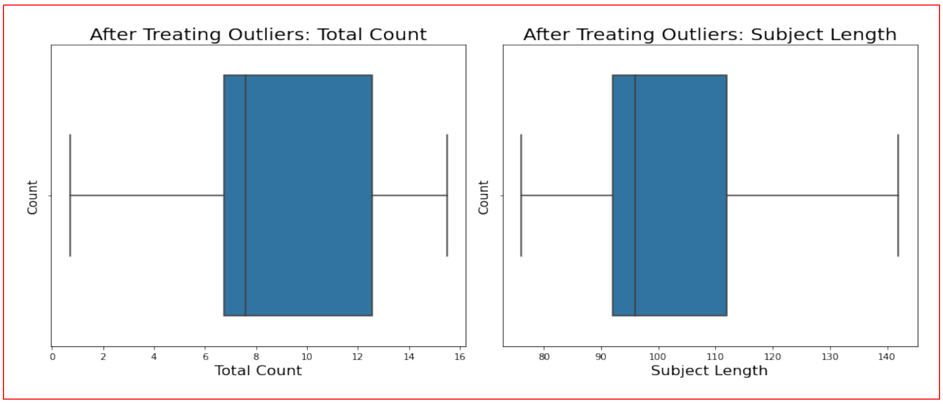

Figure 7: Boxplot: After Treating Outliers

5.2.2 Creating Recency Score

To capture the temporal aspect of customer interactions, we engineered new features calculating the time differences between when a message was sent and when it was opened, clicked, or resulted in a purchase. Next, the time differences were segmented into 10 categories (i.e., 10 to 0) using ”pd.cut” function, where higher values were given to correspond to shorter times (i.e., more recent interactions) and lower values to longer times. NaN values were filled with 0 to maintain data integrity during recency calculations, as these are the customers who didn't click the message.

Sent Date	**Clicked Date**	**Difference in Days**	**Recency Score**
01-01-2023	01-01-2023	0	10
01-01-2023	02-02-2023	32	9
01-01-2023	03-03-2023	61	8
01-01-2023	01-01-2024	365	1
01-01-2023	Did not Click	NA	0

Table 2: Recency Scores based on Clicked Date

From Table 2, we can see how the Recency Score was calculated for Message Clicked feature. Minimum the time difference between messages sent and clicked, higher the recency score. If the customer did not click the message, then the recency score would be 0. Similarly, recency scores were created for message opened and purchase.

5.2.3 One-hot Encoding to handle Categorical Variables

One-hot encoding was applied to categorical variables in the dataset to transform them into a numerical format. One-hot encoding creates binary columns for each unique category. The drop_first=True parameter is used to avoid multicollinearity by dropping the first category in each set of dummy variables. This step was important to preserve the critical information of categorical variables.

The feature engineering steps described are very critical in making the raw dataset well-structured for uplift modeling. The treating of outliers and handling of categorical features using one-hot encoding were some of the key steps that had to be addressed before the next step.

5.3 Modeling

5.3.1 Two Model with CatBoost Classifier

The Two-Model approach is a popular technique in uplift modeling. It involves training two separate models. Treatment Model (estimator_trmnt): This model is trained on the subset of data where customers received the treatment. Control Model (estimator_ctrl): This model is trained on the subset of data where customers did not receive the treatment. The difference between the predictions of these two models gives the uplift score. From this score, we can estimate the incremental effect of the treatment. In our case, it's determining whether a customer will likely click on the message or not.

Two-Model is imported from "scikit-uplift" library, which is specifically designed for uplift modeling. CatBoostClassifier is used for this model. CatBoost is a gradient boosting algorithm that handles categorical features naturally, this is used in e-commerce data where categorical features are more. The CatBoostClassifier is used for both treatment and control models. The 'ddr_control' is applied, which is a method that specifies how the difference in predictions should be calculated. The control model's predictions are directly used as the baseline to measure uplift. The Two-Model approach provides a structured approach to understanding and predicting customer responses.

5.3.2 Class Transformation with CatBoost Classifier

Class Transformation is a method in uplift modeling, where the problem is transformed into a classification task. Two datasets are created here, one for the treatment group and one for the control group. The classifier used here is CatBoost which is trained to predict a "pseudo-outcome," that calculates the difference in outcomes between treated and control groups. The pipeline is used for this code, to simplify the workflow of multiple processing steps. The pipeline is trained on the training data (X_train, y_train) with the corresponding treatment group (trmnt_train).

Stratify split is applied during the splitting of data. The stratify parameter in train_test_split ensures that the distribution of the target variable and the treatment indicator remains consistent across both the training and validation sets. This is done to prevent any data imbalances that could affect model performance. This is an important step in uplift modeling, where both the treatment and control groups need to be correctly represented.

5.3.3 Two Model with LightGBM Classifier

LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification, and many other machine learning tasks. The advantages of this algorithm are faster training speed, higher efficiency, lower memory usage, and capable of handling large-scale data.

Both treatment and control datasets are being trained on the LGBM Classifier. The n_estimators is set to 100, which controls the number of boosting rounds that the model will build. As the number of trees is increased to enhance the model performance, there is a risk of overfitting. Hence, the learning rate is set to a moderate value of 0.05. A lower learning rate reduces the impact of each individual tree, requiring more trees to achieve the same level of performance but making the model more robust and less prone to overfitting.

Both the regularization techniques are applied, ’lambda_l1’ and ’lambda_l2’. This is important as it controls the model by penalizing large coefficients. L1 regularization introduces sparsity in the model, which can be particularly beneficial when working with high-dimensional data, as it forces the model to rely on fewer, more meaningful features. L2 regularization, on the other hand, helps to prevent overfitting. Additional hyperparameter tuning is done such as ’min_split_gain’ and ’min _child_weight’. These are designed to ensure that only the most meaningful splits in the tree-building process are pursued. Additionally, ’subsample’ and ’colsample_bytree’ parameters are set to reduce the potential for overfitting, which introduces randomness and makes the model more generalizable to unseen data.

The two models initialized for treatment and control are passed as the treatment and control estimators respectively. The method parameter, method=’vanilla’ indicates that a standard Two-Model approach is being used. Here, two separate models are trained: one on the treated population and the other on the control population. The difference in predictions between these two models gives the uplift. The model is retrained multiple times with multiple hyperparameter tunings.

5.3.4 T-Learner with LightGBM

The T-learner is a specific type of meta-learner used in causal inference and uplift modeling. T-learner is similar to a two-model approach. Instead of training a single model on all the data, we train two models, on two different subsets of data.

The models are trained on LightGBM and the key parameters in this model are ’n_estimators’, ’learning_rate’, and regularization parameters such as ’lambda_l1’ and ’lambda_l2’. Learning rate is set low, which makes model more robust by ensuring that each tree makes only a small contribution to the overall prediction. Subsampling and feature sampling (’colsample_bytree’) is applied which introduces randomness into the model and increases the generalization properties. Subsample parameter is set to 0.8 which indicates that each tree will be built using 80% of the data. All these parameters are tried and set with multiple iterations to balance the complexity of the model so that overfitting cannot take place. Finally, the early stopping is applied to stop the training process when the model starts to overfit or there is no increase in performance.

Section 6 describes the evaluation of these models in detail.

6 Evaluation

The evaluation metrics used to evaluate the implemented models are Uplift Score at 30%, Area Under the Uplift Curve (AUUC), and Area Under the Qini Curve (AUQC). These metrics give insight into how well each model discriminates between treatment and control groups, identifying customers likely to respond positively to marketing intervention.

Figure 8 shows the uplift scores of all models tested in this study at 30%. The Class Transformation model with CatBoostClassifier achieved the highest uplift score of 7%. This indicates that it can strongly identify customers who would respond positively to the marketing campaigns. Meanwhile, Two-Model with both CatBoostClassifier and LGBMClassifier achieved a lower uplift score of around 0.15%, indicating lower discrimination power between treatment and control groups. T-learner with LGBM regressor performed slightly better with an uplift score of 0.32%, but this was still not good enough to show

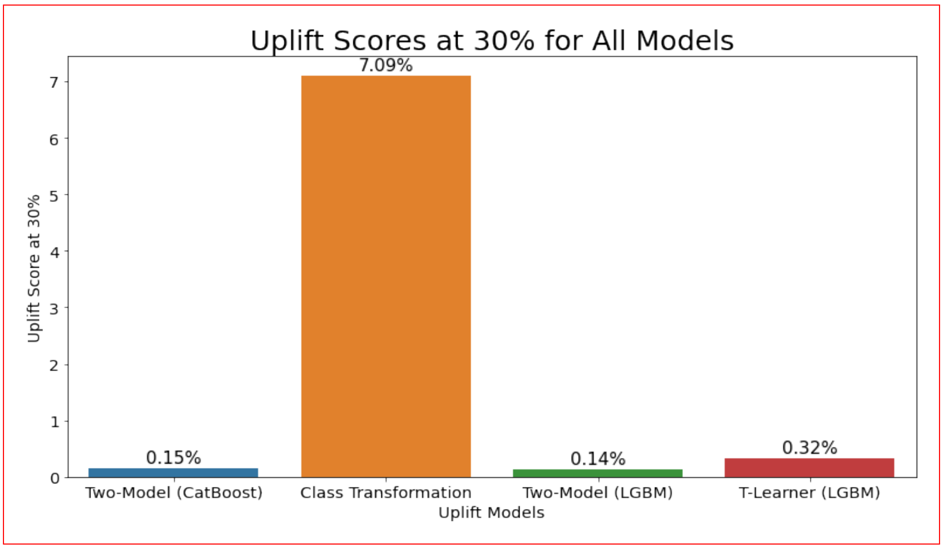

Figure 8: Uplift Score @ 30% of the models

substantial impact. These results highlight the challenges of applying uplift modeling in the fashion e-commerce domain. This is because of the complex customer behavior, which is hard to model by conventional approaches.

ID	Models	AUUC	AUQC
1	Two-Model (CatBoost)	-0.11	-0.14
2	Class Transformation (CatBoost)	0.03	-0.16
3	Two-Model (LightGBM Classifier)	-0.11	-0.14
4	T-Learner (LightGBM)	0.28	0.37

Table 3: Comparison of Uplift Models: AUUC and AUQC

Table 3 shows the comparison of Area Under the Uplift Curve (AUUC), and Area Under the Qini Curve (AUQC) across the four uplift models. T-Learner with LGBM model achieved highest AUUC and AUQC scores with 0.28 and 0.37 respectively. However, these scores are minimal. Despite Class Transformation's good uplift score, the AUUC and AUQC are low and negative respectively. The Two-Model Approach with CatBoostClassifier and LGBMClassifier both produced negative scores, reflecting a poor ability to generalize across the dataset. Among the four models, the T-Learner with LGBM Regressor showed slightly better performance with positive AUUC and AUQC scores, suggesting a modest but notable improvement over the other models.

The ideal range for both these scores in the industry is generally positive. Higher positive values imply that the model is successfully identifying customers who are more likely to respond to marketing interventions. However, the model's evaluation scores were negative, including -0.11 in the Two-Model approaches. And minimal improvement of 0.28 AUUC score by T-Learner with LGBM. The reasoning for this might be that the fashion e-commerce ecosystem is very complicated, where customer engagement and buying behavior are influenced by numerous factors. Further discussion on the performance of the model and its evaluation metrics is in Section 6.1.

6.1 Discussion

The uplift score at 30% is a crucial metric in evaluating the model performance. In this study, the Class Transformation model with CatBoost achieved the highest uplift score at 30%, with a value of 7.09%. This suggests that this model was moderately successful in targeting customers who would be influenced by the marketing messages, making it potentially useful in real-world applications. However, the Two-Model approach with CatBoost, with LGBM, and the T-Learner with LGBM produced much lower uplift scores. These lower scores indicate a limited ability to accurately identify the top 30% of responsive customers.

The AUUC and AUQC scores of the models applied in this study were low and negative. Despite hyperparameter tuning models such as applying regularization techniques, optimizers, feature sampling, and subsample to prevent overfitting. These parameters were not fully optimized to handle the specific characteristics of the data from the Russian domain in fashion e-commerce. This includes such factors as seasonal trends and regional purchasing behaviors. The results from Nandy et al. (2023), where models like the Generalized Causal Tree for Uplift Modeling demonstrated stronger performance in the e-commerce domain. These models achieved higher AUUC and AUQC scores, indicating a more effective separation between treatment and control groups. These studies usually apply advanced techniques, such as adversarial training and feature desensitization to decrease the bias induced by complex data sets. The lower scores in our study highlight the potential value in adopting more advanced neural network techniques, possibly similar to the ones used in Wang, Ye, Wang, Zhou, Zhang, Zhang, Jiang and Zhai (2024) Sun and Chen (2024), but which could not be implemented due to limitations in computational resources and further in-depth knowledge about the concepts used.

The models applied in this study struggled to generalize, possibly due to the nature of the data or the limitations of the modeling techniques themselves. As we compare the models used in this project and similar research conducted by Sun et al. (2023). It can be noted that the shortcoming in our design indeed plays a big role in the less-than-perfect outcomes. The models used in this study, the Two-Model Approach with CatBoost, LGBM, and Class Transformation could not achieve such high scores.

The study, despite the limitations and very modest results, provides some sort of contribution to the knowledge base in the field of uplift modeling for the Russian fashion e-commerce market. The findings emphasize that customer behavior in e-commerce is complex and there is a need for more robust and sophisticated modeling techniques.

7 Conclusion and Future Work

This paper aimed to explore the effectiveness of uplift modeling in the fashion e-commerce domain using a Russian online fashion retailer as a case study. Four different uplift models were trained and evaluated in this research. Two-Model Approach, Class Transformation, and T-Learner with LightGBM. The most successful one was the Class Transformation model with CatBoost Classifier, which achieved an uplift of 7% at 30%, indicating its potential in targeting potential customers. However, several important drawbacks were discovered in the study such as low AUUC and AUQC scores for most of the models. Indicating difficulties in generalization for the whole dataset. The results suggest that even if uplift modeling can provide insight into customer segmentation and targeted marketing, its application in a fashion e-commerce domain is challenging due to factors such as overfitting, computational demands, and complex customer behavior. Future research should focus on developing more advanced models and techniques like neural networks and region-specific adaptations. Additionally, scalability and computational efficiency of these models should be addressed in the future which is crucial in dynamic environments like e-commerce. This research serves as an introduction to the strengths and weaknesses of various uplift models in fashion e-commerce, paving the way for refined approaches in the future.

References

Bokelmann, B. and Lessmann, S. (2024). Improving uplift model evaluation on randomized controlled trial data, *European Journal of Operational Research* **313**(2): 691–707.

De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K. and Phan, M. (2021). Uplift modeling and its implications for b2b customer churn prediction: A segmentation-based modeling approach, *Industrial Marketing Management* **99**: 28–39.

Devriendt, F., Van Belle, J., Guns, T. and Verbeke, W. (2022). Learning to rank for uplift modeling, *IEEE Transactions on Knowledge and Data Engineering* **34**(10): 4888–4904.

Gubela, R. M. and Lessmann, S. (2021). Uplift modeling with value-driven evaluation metrics, *Decision Support Systems* **150**: 113648. Interpretable Data Science For Decision Making.

Gubela, R. M., Lessmann, S. and Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling, *European Journal of Operational Research* **283**(2): 647–661.

Ke, W., Liu, C., Shi, X., Dai, Y., Yu, P. S. and Zhu, X. (2021). Addressing exposure bias in uplift modeling for large-scale online advertising, *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1156–1161.

Kodikara, N. and Shahtahmassebi, G. (2023). Predicting potential customers in direct marketing using uplift modelling and advanced machine learning, *2023 International Conference on Computer and Applications (ICCA)*, pp. 1–6.

Michalský, F. and Kadıoğlu, S. (2021). Surrogate ground truth generation to enhance binary fairness evaluation in uplift modeling, *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1654–1659.

Nandy, P., Yu, X., Liu, W., Tu, Y., Basu, K. and Chatterjee, S. (2023). Generalized causal tree for uplift modeling, *2023 IEEE International Conference on Big Data (BigData)*, pp. 788–798.

Peralta, B., López, M., Ruiz, J., Nicolis, O. and Caro, L. (2023). Uplift modelling applied to a chilean retail company with siamese neural networks, *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pp. 1–6.

Sato, M., Kawai, S. and Nobuhara, H. (2019). Action-triggering recommenders: Uplift optimization and persuasive explanation, *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 1060–1069.

Shimizu, A., Togashi, R., Lam, A. and Huynh, N. V. (2019). Uplift modeling for cost effective coupon marketing in c-to-c e-commerce, *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1744–1748.

Singh, S. S. K., Kumar Sinha, A., Pandey, T. N. and Acharya, B. M. (2023). A machine learning approach to compare causal inference modelling strategies in the digital advertising industry, *2023 2nd International Conference on Ambient Intelligence in Health Care (ICAIHC)*, pp. 1–7.

Sun, Z. and Chen, X. (2024). M3tn: Multi-gate mixture-of-experts based multi-valued treatment network for uplift modeling, *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5065–5069.

Sun, Z., He, B., Ma, M., Tang, J., Wang, Y., Ma, C. and Liu, D. (2023). Robustness-enhanced uplift modeling with adversarial feature desensitization, *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1325–1330.

Vairetti, C., Gennaro, F. and Maldonado, S. (2024). Propensity score oversampling and matching for uplift modeling, *European Journal of Operational Research* **316**(3): 1058–1069.

Wang, D., Xu, Q., Feng, Y., Ignatius, J., Yin, Y. and Xiao, D. (2024). Uplift modeling and its implications for appointment date prediction in attended home delivery, *Decision Support Systems* **185**: 114303.

Wang, H., Ye, X., Wang, Y., Zhou, Y., Zhang, Z., Zhang, L., Jiang, J. and Zhai, Y. (2024). Uplift modeling based on graph neural network combined with causal knowledge, *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 1487–1492.

Wang, H., Ye, X., Zhang, Z. and Wang, Y. (2022). Multihead causal distilling weighting is all you need for uplift modeling, *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pp. 59–65.

Zaniewicz, and Jaroszewicz, S. (2013). Support vector machines for uplift modeling, *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 131–138.

Zhan, B., Liu, C., Li, Y. and Wu, C. (2024). Weighted doubly robust learning: An uplift modeling technique for estimating mixed treatments' effect, *Decision Support Systems* **176**: 114060.

Zhao, Y., Fang, X. and Simchi-Levi, D. (2017). A practically competitive and provably consistent algorithm for uplift modeling, *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 1171–1176.