Erasmus University Rotterdam,

Erasmus School of Economics

Master Thesis: Comparing the Performance of XGBoost and Shapley Values to Last Touch Attribution

HARNESSING MACHINE LEARNING FOR MARKETING CHANNEL ATTRIBUTION MODELING

Julia Tout – 528162

Date: 22, August, 2023

Supervisor: Professor Dr. Kathrin Gruber

Second Assessor: Dr. (Erjen) JEM van Nierop

Abstract

This research evaluates the performance of integrating XGBoost with Shapley value analysis in marketing channel attribution modeling. It contrasts this approach with Last Touch Attribution (LTA), a popular heuristic that credits conversion entirely to the last channel. Despite LTA's popularity, research shows that it may not be suitable for longer and more complex customer journeys. Such multifaceted journeys require more advanced attribution models capable of accurately distributing credit across the diverse channels used (Buhalis & Volchek, 2021). This, in turn, empowers marketers to allocate their resources strategically, which may enhance the efficiency of their marketing campaigns (Kannan et al., 2016). The central objective of this paper is to determine whether the implementation of XGBoost with Shapley values presents a viable solution to the attribution challenge. The Gaussian Process Boosting (GPBoost) algorithm, a specialized variant of XGBoost, was selected to accommodate the nature of the data used. In a comparative analysis against the Linear Mixed Effects Logistic Regression (LMER), the GPBoost model showcased exceptional performance, attaining an Area Under the ROC Curve of 0.86, notably surpassing the 0.70 achieved by LMER. Following a thorough evaluation of the models' performances, a comprehensive examination of credit attribution results was undertaken. The spectrum of models employed was the following: LTA, uniform attribution, LMER, and GPBoost with Shapley values. The distribution of credit revealed nuances among the models, with both heuristic and data-driven approaches demonstrating instances of alignment. These results, as well as the limitations of this research, were then discussed. Finally, the contribution of this study to the dynamic landscape of data-driven attribution modeling was assessed.

Keywords

Attribution modeling, Machine Learning, Last Touch Attribution, XGBoost, GPBoost, Shapley Values, SHAP, Linear Mixed Effects Logistic Regression, Uniform Attribution, Variable Importance, Explainable Machine Learning, Conversion, Touchpoint Attribution, Digital Marketing.

Contents

1 Introduction

In recent years, attribution modeling has indeed gained significant prominence in marketing due to the growing recognition of the importance of touchpoints in the customer journey. Touchpoints refer to the various interactions or marketing channels customers encounter while engaging with a business. These can include advertisements, social media engagements, website visits, and email campaigns, to name a few. The customer journey encompasses the complete process of customer interactions and touchpoints, from initial awareness to the final action taken by the customer. Accurate attribution of conversions to specific touchpoints provides valuable insights into the touchpoints that are most influential in driving customer actions. This understanding allows businesses to optimize their marketing strategies, allocate resources effectively, and enhance campaign performance to achieve a higher return on investment.

Machine learning techniques have emerged as powerful tools for attribution (Shao & Li, 2011; Dalessandro et al., 2012; Li & Kannan, 2014; Xu et al., 2014; Berman, 2015; Anderl et al., 2016*; Zhao et al., 2018; Mahboobi et al., 2018;* Kadyrov & Ignatov, 2019; Buhalis & Volchek, 2021). These techniques leverage data patterns to predict and analyze customer behaviors, providing businesses with a comprehensive understanding of the customer journey. Machine learning enables businesses to analyze customer behaviors and attribute conversions – the desired action taken by a customer, such as making a purchase – to specific touchpoints, further enhancing their understanding of the customer journey and its influence on conversions.

While machine learning approaches have gained significant traction, it is essential to acknowledge the popularity of a widely used heuristic called last touch attribution (LTA). LTA attributes customer conversions solely to the last interaction before converting. Despite its simplicity as a heuristic, it has gained widespread adoption in the industry, mainly due to its integration into the Google Analytics algorithm. Google Analytics, a widely used web analytics service provided by Google, is employed by 84.1% of the top 10 million websites worldwide (W3Techs, 2023). The availability of a free version has further contributed to its prevalence, especially among small enterprises, as it offers easy access to valuable insights into metrics such as customer conversion and top-performing products/services (Google Marketing Platform, 2023a, 2023b).

The traditional approach of last touch attribution (LTA) has received criticism, specifically when used on longer customer journeys, for oversimplifying the complex customer journey and neglecting the contributions of other touchpoints (Singal et al., 2022; Buhalis & Volchek, 2021). This approach has resulted in decreased profits for global advertisers, in some instances, compared to not using any attribution at all, as emphasized by Berman (2018). However, empirical evidence has shown that advanced marketing attribution models, including machine learning models, effectively optimize marketing return on investment (de Haan, Wiesel, & Pauwels, 2016; Kireyev et al., 2016). In light of this, researchers have increasingly conducted studies comparing machine learning-based advanced attribution models with simpler models like LTA. Their objective is to identify a model that provides a more accurate understanding of the contribution of different touchpoints throughout the customer journey. By doing so, they aim to equip businesses with the necessary tools to effectively allocate resources and optimize their marketing strategies, leading to improved outcomes and return on investment (Shao & Li, 2011; Dalessandro et al., 2012; Li & Kannan, 2014; Xu et al., 2014; Berman, 2015; Anderl et al., 2016).

In this research, the XGBoost model is chosen as the machine learning approach. XGBoost is a highly effective and widely used gradient-boosting algorithm that has demonstrated superior performance in various prediction tasks. Its ability to handle complex relationships and interactions among variables, even in the case of non-linear relationships, makes it well-suited for modeling the complex dynamics of customer conversion (Chen & Guestrin, 2016). Furthermore, to assess the contribution of different touchpoints toward customer conversion, the research will leverage Shapley values. Shapley values are a well-established concept in cooperative game theory that provides a fair and interpretable measure of each touchpoint's influence on the outcome (Lundberg & Lee, 2017). By incorporating Shapley values into the XGBoost model, the research aims to provide accurate predictions and comprehensible feature importance measures that identify the touchpoints that have the most significant impact on conversions.

This idea is further supported by various studies surrounding the topic of attribution modeling. For instance, Dalessandro et al. (2012) highlighted the ability of Shapley values to provide a more comprehensive and accurate approach to credit allocation in multichannel marketing, Li and Kannan (2014) utilized Shapley values in their nested measurement model to attribute credit to different channels, and Kadyrov and Ignatov (2019) found that their attribution model based on gradient boosting over trees outperformed bagged logistic regression and Markov chains in terms of the ROC AUC metric.

Despite the existing studies on the topic, crucial aspects still require further investigation. While some studies have examined touchpoint attribution, most have primarily focused on evaluating individual touchpoints in isolation without considering their collective impact on customer conversions. Additionally, limited research has been conducted on the comprehensibility and interpretability of feature importance measures, such as Shapley values. Furthermore, the untapped potential of harnessing XGBoost, a robust machine learning algorithm, in the attribution modeling domain remains largely unexplored. Although XGBoost has found wide application in various domains, its specific performance and application in attribution modeling have yet to receive adequate attention. Therefore, there is a compelling need to thoroughly investigate and compare the performance of an XGBoost model with Shapley value analysis against conventional last-touch attribution models in accurately assessing the contribution of different touchpoints to customer conversions.

This research aims to bridge this gap by offering valuable insights into the comprehensibility of feature importance measures provided by XGBoost models with Shapley values, providing practical recommendations for businesses to enhance their conversion prediction and optimization strategies. The primary focus is to explore the effectiveness of XGBoost models in attributing the contributions of different touchpoints towards customer conversions and compare it with conventional last-touch attribution models, ultimately improving the accuracy of attribution modeling and optimizing marketing strategies.

2 Literature Review

2.1 General overview

A recurring dilemma for marketers arises from the challenge of quantifying and comprehending the effectiveness of advertising campaigns. It is crucial to accurately evaluate which advertisements drive the most conversions as this allows marketers to optimize marketing strategies and allocate resources effectively (Kannan et al., 2016). Nevertheless, correctly attributing credit to a particular advertisement is a complex task. Advertising campaigns incorporate various overlapping marketing channels and customers typically interact with multiple channels across different devices along their journey. Each one of these marketing channels and touchpoints may play a role in the end decision to make a conversion or not, making it challenging to isolate the contributions of individual interactions. As a result, the question of how to attribute the influence of marketing channels accurately arises. If marketers don't properly attribute credit to the different channels and touchpoints, they may misallocate marketing budgets. This could

mean spending on channels that do not greatly enhance conversion rates while under-investing in channels that do. This reduces the effectiveness of the marketing strategy as a whole, resulting in a suboptimal return on investment and a missed opportunity to attract new profitable customers (De Haan, Wiesel, & Pauwels, 2016).

While the last touch attribution model is merely a simple heuristic, it remains highly prevalent; this is partially due to the notoriety of Google Analytics, which utilizes such models. It is simple to use and highly accessible, making it an attractive option for small businesses in particular. Moreover, research has shown that it is effective when used on short customer journeys (Buhalis & Volchek, 2021). Nonetheless, heuristic models such as LTA, struggle to provide a unified view of the customer's journey. This leads to incomplete attribution and inaccurate credit allocation, specifically in longer customer journeys involving multiple touchpoints (Buhalis & Volchek, 2021). According to Thakurani (2022), this may arise because multiple touchpoints might have influenced a consumer's decision before they make a purchase. Still, the credit for the conversion is often assigned solely to the last touchpoint. As a result, the value of additional touchpoints may be overlooked or drastically undervalued.

Cui, Ghose, Halaburda, Iyengar, Pauwels, Sriram, Tucker, and Venkataraman (2021) elaborate that there is a growing tendency for customers to switch between both devices and marketing channels throughout their customer journey, further exacerbating this dilemma. While this is a challenge for attribution modeling generally, conventional heuristics reliant on tracking touchpoint sequences across devices are especially impacted. The underlying data in such cases is intricate and frequently scattered, complicating the task of presenting a cohesive picture of the customer's interactions. Subsequently, attributing the final conversion to a single touchpoint becomes risky and inadequate. Since conventional first or last touch models overlook these intermediate touchpoints, their contribution is incompletely assessed (Cui et al., 2021). Certain touchpoints may serve as initial points of awareness, while others play a role in the final decision-making process, yet only the latter would be considered in these models. Nonetheless, these earlier awareness-building channels crucial in generating conversions could be disregarded in the attribution process. Instead, channels that happened to be the final touchpoint before conversion will be prioritized. This ultimately could lead to inefficient allocations of marketing budgets (Thakurani, 2022).

This can also be seen as a failure of these traditional models to take into account the sequential effects of touchpoints, where the order of interactions can influence customer behavior (Cui et al., 2021). The order of interactions and time lags between touchpoints can significantly

influence customer actions, and models that overlook these temporal dynamics may misallocate credit and underestimate the true impact of each touchpoint. Furthermore, traditional heuristics neglect the synergy effects of multiple touchpoints working together to drive conversions. By treating touchpoints as individual entities, these models can undervalue certain channels and miss out on optimization opportunities. Nonetheless, these limitations extend to more complex models, including traditional Shapley values, where sequential and time-related effects are similarly overlooked.

Cui et al. (2021) propose remedies to overcome the aforementioned attribution challenges in omnichannel marketing. They first stress the importance of replacing traditional single-touch attribution models with more sophisticated approaches, such as multi-touch or algorithmic models. They also emphasize the need for improved tracking capabilities. The papers by Kannan, Reinartz, and Verhoef (2016) and Buhalis and Volchek (2021) likewise illustrate the consensus amongst the literature advocating for the creation of advanced attribution models that are capable of considering numerous customer touchpoints. Buhalis and Volchek (2021) argue that simple multi-touch attribution, also known as fractional attribution, may not fully capture realistic consumer behavior either. In contrast, incremental attribution methods theoretically assign value to each touchpoint while considering the cumulative effect between them, making them a more realistic approach. However, more research needs to be done on these to be able to create a viable and standardizable framework.

The marketing attribution taxonomy by Buhalis and Volchek (2021) and the recommendations proposed by Cui et al. (2021) emphasize the importance of leveraging data-driven approaches and advanced analytics to develop more accurate attribution models that consider the impacts of sequence, timing, and synergy effects among channels. According to Buhalis and Volchek (2021), custom attribution models, such as algorithmic and data-driven models like Markov Models or Shapley values, provide a more realistic view of the customer journey by incorporating all events and accounting for customer heterogeneity. They highlight the advantage of data-driven attribution in accurately determining touchpoint effectiveness, especially in long customer journeys, through the use of individual-level data. This supports the general movement and extensive research advocating for sophisticated models that incorporate vast customer data and leverage big data analytics. These models provide more nuanced insights into customer behavior, as opposed to manual rule-based attribution methods like LTA.

In a similar vein, the model by Dalessandro et al. (2015) innovates attribution modeling in a data-driven manner by incorporating causal relationships between touchpoints and

conversions. Their causal inference framework goes beyond simple correlations and investigates the true cause-and-effect relationships. By employing propensity score matching, they estimate the causal effects of individual touchpoints on conversion probabilities more accurately. Their meticulous models consider customer characteristics, touchpoint sequences, and order, providing a thorough depiction of the attribution process. This leads to more precise touchpoint effect estimation, enhanced credit allocation, and the development of robust marketing strategy models.

Anderl, Becker, von Wangenheim, and Schumann (2016) additionally make use of machine learning and advanced analytics by means of a Markov-based attribution model. Here, the customer journey is represented as a graph, in which touchpoints are nodes and the connections between them represent the sequence of customer interactions. This graph-based representation allows for a more comprehensive customer journey analysis by taking into account the order and relationships between touchpoints.

Countless more sophisticated, data-driven, analytical models have been proposed in the field of attribution modeling, emphasizing the gravity of accurately attributing credit to marketing channels. These methods facilitate the exploration of data patterns and interconnections, resulting in an improved understanding of customer behavior. This enhances the efficacy of campaigns and augments insights into customer dynamics (Kannan et al., 2016).

2.2 Shapley values

Shapley Values are originally derived from cooperative game theory. They depict a more equitable manner in which one can allocate the contribution of individual players in a coalitional game. By assessing how each player's addition impacts various subsets of the coalition, Shapley Values offer a more impartial method of distributing the collective value created by the coalition. This is done by means of evaluating the difference each player makes when added to different subsets of the coalition. In the calculation of these values, all possible permutations of players are taken incorporated, further fostering an unbiased allocation (Shapley, 1953)

In the realm of marketing attribution, Shapley values can be applied to allocate credit to various marketing channels or touchpoints: the “reward” is represented by the customer conversions, whilst the “players” are the diverse marketing channels/touchpoints impacting these conversions. Similar to the original idea of Shapley values, in marketing attribution, the contributions of these channels are quantified in the different potential combinations and permutations. Consequently, Shapley values provide a way to assess the effectiveness of marketing channels in generating customer conversions. Furthermore, the versatility of Shapley

values allows for their application across a miscellaneous array of marketing scenarios, including distinct industries, campaign settings, and touchpoint configurations. Thus, they represent a robust framework that is adaptable to the unique attributes of every marketing campaign. As such, marketers can leverage Shapley values to evaluate channel effectiveness and identify under and over-capitalized channels. Subsequently, these insights enable informed and improved decision-making on resource allocation and campaign optimization.

Nisar and Yeung (2015) used a Shapley Value regression method to determine the exact contributions and statistical significance of each explanatory variable to the variance of the dependent variable in a regression model. The Shapley Value approach takes into account the potential correlation among regressors and measures the contribution of each attribute by the improvement in R-square. They concluded that Shapley Value-based regression provided insights into the complexity of attribution modeling and demonstrated its efficacy in assigning fair rewards to multiple advertising channels.

Berman (2018) discusses the application of Shapley values in attribution in the context of online advertising. He discusses two desirable properties of Shapley values: efficiency and marginality. These ensure that all value generated by marketing channels is allocated and that the sum of allocations matches the total revenue. Notably, Shapley value treats channels equally, regardless of their order, considering their contributions as equal. Berman's research reveals that LTA encourages excessive ad exposure, surpassing the optimal advertising level for global advertisers. In contrast, the Shapley value method adjusts ad exposure based on platform popularity, leading to increased profits. However, LTA can in certain cases enhance market efficiency beyond what can be achieved with complete market information or a comprehensive understanding of all marketing channel impacts. In these cases, the Shapley value method may decrease market efficiency.

This phenomenon develops from the complex interplay between ad allocation efficiency, measurement accuracy, and the financial gain of advertisers and publishers. The research also emphasizes that attribution may not guarantee optimal profit for advertisers; without attribution, advertisers may face less competition for the same channels, leading to lower costs and potentially higher profits for the parties involved. Alternatively, attribution may enhance global advertisers' profits, increase market efficiency, and potentially benefit publishers. Therefore, attribution remains a viable alternative for advertisers to optimize their strategies in the online advertising market.

Zhao, Mahboobi, and Bagheri (2018) highlight the advantages of using Shapley values in attribution modeling for multichannel digital marketing campaigns. They argue that Shapley values provide a transparent and objective, yet computationally rigorous approach to assigning credit to advertising channels. It considers channels' interconnectedness while maintaining an unbiased evaluation process. The impartiality inherent in Shapley values enables them to effectively capture the marginal contribution of each channel. This empowers marketers to comprehend the influence of diverse channels and, in turn, make well-informed decisions regarding resource allocation. Shapley values are consistent and stable, ensuring coherent representation and consistent attribution across various channel permutations.

This is in line with the research done by Mahboobi, Usta, and Bagheri (2018), who found that implementing Shapley values at a large scale using logistic regression was effective in attribution modeling for real online campaigns. They emphasized the benefits of Shapley values in understanding the contribution of different marketing inputs to conversions. Nevertheless, the authors of both articles also agree that calculating Shapley values is a computationally expensive task, especially with large datasets. To solve this issue, Zhao, Mahboobi, and Bagheri (2018) propose a model that calculates Shapley values more resourcefully, while Mahboobi, Usta, and Bagheri (2018) suggest using hierarchical modeling and probabilistic approximations to improve computational efficiency.

The research by Singal, Besbes, Desir, Goyal, and Iyengar (2022) explores the application of Shapley Value as a metric for attribution in the context of a Markov model. They identify that using Shapley values for attribution modeling in their given context presents a drawback of not adjusting for counterfactuals; these are alternative paths or actions that could have been taken by customers but were not. Sharma, Li, and Jiao (2022) also mention the same drawback in their research paper. Therefore, both papers propose and provide theoretical and empirical evidence that supports a counterfactual adjusted Shapley Value metric for attribution, which aims to capture the contributions of past actions while incorporating counterfactual reasoning.

In contrast, Huang and Marques-Silva (2023) highlight in their research that existing approaches approximating Shapley values for feature attribution in explainability potentially incorrectly attribute features' importance. They demonstrate several situations in which Shapley values do not correlate with feature relevancy: (i) irrelevant features can display non-zero Shapley values, (ii) Shapley values may result in irrelevant features ending up being seen as more important than relevant ones, (iii) relevant features may have a Shapley value of 0, depicting no importance, and (iv) pairs of features can exhibit conflicting Shapley values where one relevant

feature is considered unimportant while an irrelevant feature is allotted some importance. As such, based on their research, Shapley values may potentially not reliably reflect the actual relevancy of features for classifiers' predictions. Subsequently, given their widespread application as explainability methods, they might have to be used with caution by decision-makers as they may mislead assessments. The authors then proceed to propose an alternative measure of feature importance that respects feature relevancy and may be computationally efficient in certain settings, such as decision trees with contrastive explanations.

Nevertheless, the paper by Balkanski and Singer (2015) discusses the notion of fairness in general. The authors show that fairness can actually sometimes result in suboptimal outcomes when compared to unfair solutions, suggesting that achieving fairness may not always be feasible or optimal. The end performance of fair attribution mechanisms will depend on the specific utility function used. Here, the paper proposes a framework for optimization in procurement settings, focusing on mechanisms using Shapley values for fairness in payments. It analyzes the trade-off between fairness and optimal utility, acknowledging the challenges and limitations of achieving fairness in procurement optimization. As such, it is difficult to berate these models simply on the notion of their theoretical fairness, as fairness itself may not be the optimal judge for the most optimal model. Instead, the applicability and performance should guide the marketers on which models are best to use for their particular context.

2.3 XGBoost

Kadyrov and Ignatov (2019) explore various machine learning techniques to address the challenges of multichannel attribution, namely, Bagged Logistic Regression, Hidden Markov Chains, Shapley value Analysis, Survival Analysis, relative weights, probabilistic approaches, and Gradient Boosting over Trees (XGBoost). The authors specify that XGBoost is particularly well-suited for this type of study, given the typically low conversion rate (0.5%-2%) and highly unbalanced classes. XGBoost is proven to handle such scenarios effectively. Logistic regression is chosen for its interpretability and the bagging technique aids in addressing multicollinearity issues among independent variables. The Markov Chain model was chosen to capture the relationships between channel interactions through transitional probabilities. By selectively removing channels from the model and observing changes in conversion probabilities, they estimated the contribution of each channel. They also employed a Shapley Value Approach, which assigns specific values to individual advertising channels based on conditional expectations. They combined this with traditional Shapley values from Cooperative Game Theory to measure the contribution and importance of each channel within the multichannel attribution

problem. Based on their outcomes, the XGBoost method consistently leads to higher AUC ROC scores compared to the other methods across all three advertising campaigns tested by the researchers. Kadyrov and Ignatov (2019) also tested a meta-algorithm in which they added additional user features in combination with the XGBoost model; this model significantly improved the performance of the second advertising campaign and surpassed the other techniques in terms of AUC ROC values. Their findings suggest that XGBoost can be a valuable method to resolve the touchpoint attribution dilemma in customer conversion.

Bryan Gregory (2018) explored the effectiveness of the XGBoost algorithm in predicting customer churn using temporal data. The study demonstrated that XGBoost outperformed logistic regression and decision trees in terms of accuracy. He also conducted a feature importance analysis to identify the key factors contributing to the model's performance. While these findings do not specifically contribute to the field of attribution modeling, they display the strength of using XGBoost in predicting customer churn, which will be explored in this research.

While research on XGBoost in attribution modeling, especially when combined with Shapley value analysis and compared to last-touch attribution (LTA), is limited, it presents an intriguing area for exploration. The existing literature suggests that Shapley value analysis works well for credit attribution, but its performance can be further enhanced when combined with a machine learning approach. Moreover, the research by Lundberg, Erion, and Lee (2019) reinforces the usage of Shapley values alongside tree ensemble methods such as XGBoost, as they discovered that it leads to improved and more consistent feature attributions. This, alongside the consensus in favor of data-driven methods over heuristic approaches, illustrates that applying XGBoost alongside Shapley value analysis could be a promising attribution modeling approach.

3 Methodology

3.1 Last-Touch Attribution Modeling

Last touch attribution is often used as a heuristic to determine which marketing channel led a customer to convert. It assigns all the credit for the conversion to the last channel a customer utilized before they made the conversion. Therefore, it assumes that the final touchpoint most heavily influences the conversion decision of a customer. Since only the last touchpoint is considered, it is a straightforward and intuitive method, making it an easy approach to use and comprehend. The input of this model will simply be the last marketing channel/touchpoint used by a customer before converting (conversion = 1), meaning that the rows with the preceding marketing channels used will be disregarded.

3.2 XGBoost

XGBoost, classical boosting, and random forest are all widely-used ensemble learning methods that use an amalgamation of models to ameliorate prediction accuracy. That being said, they each make use of distinct approaches and principles to reach their outcomes.

Classical boosting algorithms, such as AdaBoost, sequentially train a series of weak learners – which are simple models that perform slightly better than random guessing – on the same dataset. These weak learners are normally decision stumps, which are basic decision trees made up of only one split. This simplicity allows them to capture basic patterns in the data and contribute to the overall ensemble model (Freund & Schapire, 1997).

Suppose we have a training dataset X, where x_i represents a training example, with corresponding binary labels Y, where y_i takes on the values 0 or 1. The first step is to then initialize the example weights w_i for each of the training examples, where i ranges from 1 to n. These weights are originally all equal to one another, taking on a value of $1/n$ (n being the total number of observations). We then iterate over the weak learners, which are denoted by $h_t(x)$, where *t* represents the iteration and *x* is the input data. During each iteration, a weak learner is trained on the dataset X using the example weights w_i. After this, the performance of each weak learner in capturing patterns in the data is evaluated by calculating the weighted error ε_t at the iteration *t*. This is done using the following formula:

$$\varepsilon_\mathrm{t} = \sum_i((w_i) \times \mathrm{I}(\mathrm{y}_i \neq h_t(x_i))) \tag{1}$$

$\mathrm{I}(\mathrm{y}_i \neq h_t(x_i))$ represents an indicator function that is equal to 1 if the prediction of $h_t(x_i)$ does not match the actual value of y_i, and equal to 0 if it does. Using this value, the weight α_t assigned to $h_t(x)$ is calculated. α_t is different from the aforementioned example weight w_i, as it represents the contribution of the weak learner to the final prediction, while w_i represents the influence of the training example. As such, α_t is obtained through the following formula:

$$\alpha_t = \frac{1}{2}(ln(\frac{1-\varepsilon_t}{\varepsilon_t}) \tag{2}$$

One can see that a lower weighted error ultimately results in a higher weight for the weak learner; this subsequently means that it does a better job at classifying the samples. Upon acquiring this value, we can then update the original example weights w_i using the following formula:

$$w_i = w_i \times e^{(-\alpha_t \times y_i \times h_t(x_i))} \quad (3)$$

As said, this weight update emphasizes the misclassified samples from the prior iterations by incrementing their respective weights. This increases the influence they have in succeeding iterations. This is particularly achieved through the exponential function, which amplifies the weight change for these misclassified samples. Subsequently, the example weights are normalized to guarantee that their relative proportions are maintained and that they add up to 1. This is done by simply dividing each weight by the sum of all example weights. At last, the weak learners' predictions are amalgamated in order to create the strong classifier $H(x)$. This is done through the following computation:

$$H(x) = \begin{cases} 1 \; if \; \sum_t \alpha_t \times h_t(x) > 0 \\ 0 \; if \; \sum_t \alpha_t \times h_t(x) < 0 \end{cases} \quad (4)$$

Random forest algorithms take a different approach in which an ensemble of decision trees is constructed independently. Using bootstrap sampling (creating a random subset with replacement), each tree is trained on a different random subset of the training data; at each split, only a random subset of features is considered. The number of features considered at this stage can be tuned when constructing the random forest to optimize the performance of the model. Nonetheless, it is often taken to be the square root of the total number of features. The randomness created by bootstrapping establishes diversity between the trees created and reduces the risk of overfitting to the training data. To evaluate the splits at each node as well as the root node, a measure of impurity called the Gini Index is often utilized. This is calculated with the following formula:

$$GiniIndex = 1 - (p^2 + (1-p)^2) \quad (5)$$

This measure quantifies the impurity at each node by considering the proportions of positive and negative examples (i.e. conversions and non-conversions, respectively). In the formula, p represents the positive examples. Selecting the splits that minimize the Gini Index allows Random Forests to construct decision trees that aim to maximize the separation between the two classes. The end prediction is then performed by aggregating the predictions of all the trees, usually through a majority vote or by taking the average (Breiman, 2001).

Extreme Gradient Boosting (XGBoost) combines the strengths of both classical boosting and random forest while introducing several improvements. It essentially is a regression and classification method renowned for its high accuracy, scalability, and efficiency. Initially, the ensemble model is empty; hence, an ensemble of decision trees is first iteratively trained, with

each new tree attempting to fix the mistakes made by the one before it. This process begins by making an initial prediction p_i^0, which is often taken as the average of the binary labels – 0.5 in this case. The discrepancy between the target variable's actual and anticipated values is known as the residual error R. One can then quantify the performance of the prediction using a loss function. For classification applications in XGBoost, the loss function is the negative log-likelihood, depicted in the following formula:

$$L(y_i, \mathrm{p_i}) = -\left[y_i\left(log(p_i)\right) + (1 - y_i)\left(log(1 - p_i)\right)\right] \quad (6)$$

Here, y_i refers to the y-axis values for one of the observed values, which can be either 0 or 1 (non-conversion vs. conversion, respectively) and $\mathrm{p_i}$ is the predicted value, between and including 0 and 1. XGBoost uses this loss function to then build its trees by minimizing the following equation:

$$\left[\sum_{i=1}^{n} L(y_i, \mathrm{p_i})\right] + \gamma T + \frac{1}{2}\left(\gamma \times O_{value}^2\right) \quad (7)$$

The term γT will be looked over, as it is simply an additional parameter that encourages pruning and does not play a role in the derivation of optimal output values or similarity scores. The first part of the equation consists of the aforementioned loss function and the ending part is the regularization term (Starmer, 2020). This is a unique feature seen in XGBoost that is not present in traditional gradient boosting methods: its ability to both handle missing data and execute LASSO and Ridge regularization. These prevent the model from becoming too complex and lower the risk of the model overfitting to the training data, which would make it less generalizable by penalizing large coefficients whilst favoring smaller counterparts (Chen & Guestrin, 2016).

The objective now is to obtain an output value (O_{value}) for the leaf that will minimize the entire equation (Starmer, 2020). Since we are still optimizing the output value from the first tree at this stage, this formula can be replaced by the following:

$$\left[\sum_{i=1}^{n} L\left(y_i, p_i^0 + O_{value}\right)\right] + \frac{1}{2}\left(\gamma \times O_{value}^2\right) \quad (8)$$

Here, p_i^0 is the initial prediction made and O_{value} is the output value from the new tree. The algorithm then tries different values for this O_{value} and chooses the value that minimizes the equation's outcome. As γ increases, the optimal O_{value} decreases further towards zero, demonstrating the regularization capacity of the penalty term in the formula (Starmer, 2020). To

find the output value more easily, the second-order Taylor series approximation below is normally utilized:

$$L(y_i, \mathrm{p_i} + O_{value}) \approx L(y_i, \mathrm{p_i}) + \left[\frac{d}{dp_i} L(y_i, \mathrm{p_i})\right] O_{value} + \frac{1}{2}\left[\frac{d^2}{dp_i^2} L(y_i, \mathrm{p_i})\right] O_{value}^2 \quad (9)$$

This equation is simply made up of the loss function of the previous prediction as well as the first and second-order derivatives of this loss function. The first-order derivative can then be represented by the Gradient of the loss function, g, while the second-order can be represented by the Hessian, h (Starmer, 2020). This will simplify the equation to the following:

$$L(y_i, \mathrm{p_i} + O_{value}) \approx L(y_i, \mathrm{p_i}) + g \times O_{value} + \frac{1}{2} \times h \times O_{value}^2 \quad (10)$$

Taking the derivative of the expansion of this function with respect to the Output value, setting it to zero, and solving for the output results in the following:

$$O_{value} = \frac{-(g_1 + g_2 + \cdots + g_n)}{(h_1 + h_2 + \cdots + h_n + \gamma)} \quad (11)$$

Deriving the first and second derivatives of equation 6, we obtain $g_i = -(y_i - p_i)$and $h_i = p_i \times (1 - p_i)$. The numerator will therefore then be equivalent to the sum of the residuals. Replacing these values in the previous formula, the output value becomes equal to the following:

$$O_{value} = \frac{\sum Residual_i}{\sum(previous\ probability_i \times (1 - previous\ probability_i)) + \gamma} \quad (12)$$

The next step now becomes computing the similarity score, which measures the similarities between the predicted values and the true values of the target variable. To do so, the optimal output value is plugged into equation 10 and the equation is multiplied by -1, such that the output value is represented by the maximum point on the curve (Starmer, 2020). This will then result in the following equation:

$$Similarity\ Score = \frac{(\sum Residual_i)^2}{\sum(previous\ probability_i \times (1 - previous\ probability_i)) + \gamma} \quad (13)$$

In summary, in XGBoost, the performance of the already-existing trees is evaluated in an effort to improve the outcomes every time a new iteration is introduced. This continues until the model's performance does not further improve or until the desired number of trees is attained (Chen & Guestrin, 2016). Therefore, gradient boosting is often used to generate an ensemble with strong prediction power when unanticipated data is incorporated. Using a gradient descent approach, the loss function, which measures the residual error, is minimized throughout the

XGBoost model's optimization stage by repeatedly changing the weights of the decision trees. In doing so, it sets the decision trees' initial weights before readjusting them by moving toward the loss function's negative gradient until the training set's smallest loss is obtained (Chen & Guestrin, 2016).

Moreover, to accelerate the model training process, XGBoost utilizes parallelization and tree pruning. It also uses distributed computation and sparse input formats, making it favorable to utilize when working with large datasets. By tuning hyperparameters such as the learning rate, maximum depth of trees, and regularization intensity, the performance of the model can be improved; cross-validation and grid-search are two methods that may be used for this (Chen & Guestrin, 2016).

Overall, in comparison to other ensemble attribution methods, such as Random Forest and AdaBoost, XGBoost exhibits several advantages, leading to it ultimately being chosen for this research. While Random Forest combines the predictions of multiple decision trees, it lacks the iterative training process and specific handling of missing data and regularization techniques found in XGBoost. AdaBoost assigns higher weights to misclassified samples to improve subsequent iterations, yet differs from XGBoost in its training process and treatment of missing data and regularization (Breiman, 2001; Freund & Schapire, 1997).

3.3 Shapley Values

In the context of generating predictions, we can perceive each characteristic value of an observation as a participant in a game, where the prediction outcome serves as the reward. The Shapley value technique, derived from cooperative game theory, provides a fair and systematic approach to distributing this reward among the various features (Lundberg & Lee, 2017). This concept can also be applied to conversion attribution, where the objective is to attribute the end conversion to the marketing channels or touchpoints.

In this case, the conversion outcome represents the reward, and measuring the contribution of each feature is equivalent to attributing the conversion to the relevant marketing channels. By utilizing the Shapley value approach, we can determine the relative importance of each feature in influencing the prediction outcome and, consequently, the conversion (Berman, 2018). Similarly, this concept extends to predictions generated by the XGBoost model, enabling us to assess the impact and importance of each feature on the prediction outcome. To obtain numerical values for feature contributions, Shapley values consider all possible combinations of features and their effects. To compute Shapley values, the model's prediction is initially computed

using the input data. Subsequently, the prediction is recalculated by removing each characteristic from the input one at a time. The difference between the original forecast and the prediction after excluding a specific feature is then used to determine the contribution of that feature (Lundberg & Lee, 2017). This can be mathematically represented by the following equation:

$$\emptyset_i(v) = \sum_{S \subseteq N} \frac{|S|!(n-|S|-1)!}{n!} \big(v(S \cup \{i\}) - v(S)\big) \quad (14)$$

In the above formula, N represents the set of marketing channels, and n the number of channels. These channels are considered the players in the marketing attribution model. The coalition, denoted as S, refers to a specific subset of players or channels that work together. The cardinality of the coalition, denoted by $|S|$, represents the number of channels in the subsets S of n, excluding i. The characteristic function $v(S)$ gives the weight of each channel after calculation. The weight is then represented by $\frac{|S|!(n-|S|-1)!}{n!}$. The marginal contribution $\big(v(S \cup \{i\}) - v(S)\big)$ of a channel is computed by comparing the value of the coalition with the channel to the value of the coalition without the channel. This incremental weighted sum captures the channel's specific impact on the coalition (Zhao et al., 2018).

Tree SHAP is frequently used as an algorithm to estimate Shapley values, as it is faster and more accurate than other algorithms, particularly when the dataset possesses high dimensionality (Lundberg & Lee, 2019). This enables efficient model development and experimentation by speeding up the computation of Shapley values, allowing for quick iterations and exploration of different configurations. Secondly, faster computation facilitates (near) real-time applications where timely decision-making and personalized user experiences are crucial. It also enhances interpretability and transparency by providing rapid insights into feature contributions and a deeper understanding of the model's behavior and data patterns *(Mahboobi et al., 2018; Zhao et al., 2018)*. This speed is particularly relevant in marketing contexts where decision-making needs to be agile and data-driven. It allows marketers to identify the most effective channels and allocate resources accordingly, optimizing their marketing strategies in near real-time. Nonetheless, it is important to note that SHAP does not exactly match the exact Shapley values (Huang & Marques-Silva, 2023).

Additionally, Tree SHAP is used to show the Shapley values for specific occurrences, which adds clarity and aids in the interpretation of the results. While Shapley values can theoretically take negative values, their interpretation and application in marketing attribution or feature importance analysis often focus on the positive contributions of channels or features

towards desired outcomes, such as conversions. This is because it is commonly assumed that channels have a positive impact on the conversion probability, so negative impacts are typically not considered. The XGBoost method has characteristics similar to that of the Tree SHAP algorithm in that it adds and removes features from a particular decision tree to determine the contribution of each feature. Instead of estimating each interacting feature's contribution separately, Tree SHAP evaluates the decision tree for each instance and then calculates the feature's contribution as the weighted sum of the differences between the predicted values with the feature present and when the feature is absent. This makes the Tree SHAP algorithm more effective and scalable (Lundberg & Lee, 2019). It employs a bottom-up methodology where the values of the terminal nodes are computed first. These are subsequently combined to provide the Shapley values of the parent nodes. Thus, Shapley values aid in understanding the underlying links and interactions between features and how they affect the results.

Shapley values provide a valuable tool for conversion and touchpoint attribution by identifying the most effective channel in driving customer conversions (Berman, 2018). By assessing the contributions of each channel to the results and model performance, marketers can optimize their marketing strategies and allocate resources more effectively. If positive interaction effects are observed between channels, it indicates that combining channels can lead to improved conversion rates. This makes Shapley values a valuable blueprint for marketers to guide their resource allocation decisions and optimize their marketing strategies.

4 Data

The dataset utilized for this analysis is digital marketing panel data. It is comprised of 586,000 marketing touch-points collected during July 2018, representing a diverse set of 240,000 unique customers who generated approximately 18,000 conversions.

The dataset offers the following key features:

- Cookie: This feature serves as an anonymous customer ID. This allows for effective tracking and analysis of the progression a customer makes throughout their interactions with the company’s marketing channels.
- Timestamp: This represents the date and time of each customer's interactions.
- Interaction: A categorical variable that signifies the type of customer interaction (impression vs. conversion).

- Conversion: A binary variable indicating whether a conversion occurred or not. This is equal to 1 when a conversion is made and 0 if not.
- Channel: Specifies the marketing channel through which the customer arrived at the website. In this dataset, there are 5 potential channels: Facebook, Instagram, Online Display, Online Video, and Paid Search.

5 Implementation

The data integrated into this analysis is binary panel data. This data constitutes of hierarchical structure that features random effects stemming from the "Cookie" variable. Since it is common for a user to interact with multiple marketing channels during their journey, one user may generate multiple entries in the dataset under the same "Cookie" identifier. These entries are documented in the sequence in which the customer encountered the distinct channels, revealing valuable insights into their decision-making process. Consequently, it is important to acknowledge and accommodate these interdependencies during the construction of the machine learning models.

Addressing these intrinsic dependencies, or "random effects," is pivotal as they may influence the model's predictive accuracy and generalization capability. Appropriately addressing random effects can be done through appropriate modeling techniques, such as mixed-effects models or Gaussian process-based approaches. These allow for a better capturing of the underlying structure of the data, enabling improved predictive performance and robust inference.

To account for these dependencies, a Linear Mixed Effects Logistic Regression can be utilized within the lme4 package in R. In addition, a regular XGBoost model would not be suitable in this case. Thus, a GPBoost model was employed, using the GPBoost package in R. This model both incorporates the regular characteristics seen in XGBoost and addresses the random effects present (Sigrist, 2023). As such, both of these approaches are appropriate to use for the panel data utilized.

Another obstacle encountered in the data was the significant class imbalance in the conversions, since only 3% of the observations result in conversions. This imbalance arises partly from the data's nature, as only the last channel used before the conversion will obtain a conversion outcome of 1. All the preceding channels pertaining to that cookie will be considered impressions, obtaining a conversion outcome of 0. Addressing this class imbalance proved to be challenging and various approaches were attempted.

Adjusting the class weights allocated to each conversion outcome is one viable method to utilize here. It was not possible to adjust the class weights in the version of GPBoost that was utilized here, yet it seems as though this feature will be added soon. However, it can prove to be difficult to find the optimal values to choose for these weights in any case, as the model can be very sensitive to even minute adjustments. Even a small favoring of the minority class can make the model predict all the observations to have a conversion outcome of 1. Moreover, the computational expensiveness of the models may deem it challenging to test an array of values. That being said, a viable approach here would be to find the optimal class weights using a grid search.

An alternative, albeit less optimal, option is under-sampling the majority class (conversion = 0). To alleviate some limitations of down-sampling, a loop was executed over the machine learning models, selecting an equal number of converted and non-converted sequences and iterating this 5 times. As such, this results in 5 different LMER and XGBoost models, each created from distinct training samples. Each of these models was then evaluated on their performance using metrics such as recall, accuracy, ROC-AUC, and F1 score. The best-performing model was then selected. Moreover, a large number of cookies in the data had only used one channel and did not result in conversion. Since these do not add much insight to the analysis, 95% of these observations were removed, contributing to the class balancing process.

Other aspects considered included ensuring that the distribution of journey lengths in the training and testing sets was relatively balanced to guarantee that the models are evaluated fairly. In addition, measures were taken to make sure that Cookies with the same ID were kept together during the sampling processes and the various loops used; this allowed the model to capture the customer journey as a whole and prevented information loss due to data fragmentation.

6 Results

An amalgamation of heuristic and machine learning-based models were harnessed in order to facilitate a comprehensive analysis of their respective output. In this section, the output of the models employed will be documented. The models to be examined are last touch attribution, uniform attribution, linear mixed effects logistic regression, and the Gaussian Process Boosting algorithm as the XGBoost model.

Since LTA and uniform distribution are heuristics, their performance is difficult to assess against the data-driven models through evaluation metrics. Therefore, their outputs will simply be

compared to those generated by the data-driven methods. A visual overview of the different models' channel credit attribution can be found in Table 1 at the end of this section to facilitate this comparison.

6.1 Last Touch Attribution

The LTA model determined that Facebook was the channel generating the most conversions, with 30.1% of all conversions being allocated to it. Paid search came second, attaining 25.8%. This was followed by the Online video channel with 19.3%, then the Instagram channel with 12.7%, and at last the Online Display channel with 12.1%.

6.2 Uniform Attribution

Interestingly, the results obtained from the uniform attribution resembled the aforementioned LTA credit distribution closely. Similar to the LTA results, Facebook was the channel that drove the most conversions, with Paid search coming in second, receiving 29.6% and 26.5% of the credit allocation, respectively. This is followed by Online Video, with 19.0%, Instagram with 12.8%, and Online Display with 12.0%.

6.3 Linear Mixed Effects Logistic Regression

The LMER model is evaluated on its performance in classifying the binary outcome variable: conversion. The model's accuracy was circa 65%, signifying the proportion of the instances it predicted correctly. The ratio of true positive predictions to the total amount of positive predictions, also known as precision, was 0.675. In other words, if the model classifies an observation into class 1, it is correct 67.5% of all instances. The recall, or sensitivity of the model was 0.574. This means that the LMER model was able to determine 57.4% of the actual positive instances. The F1 obtained was 0.621; this metric serves as a balance between precision and recall in terms of measuring accuracy in both classes. Finally, the AUC, or the Area Under the (ROC) curve signifies how well the model can discriminate between the two classes of the outcome variable. The closer this value is to 1, the higher the model's discrimination capabilities are. Here, the AUC was circa 70.0. The respective ROC (Receiver Operating Characteristic) Curve, which is a graphical representation of the AUC, can be seen in the figure below.

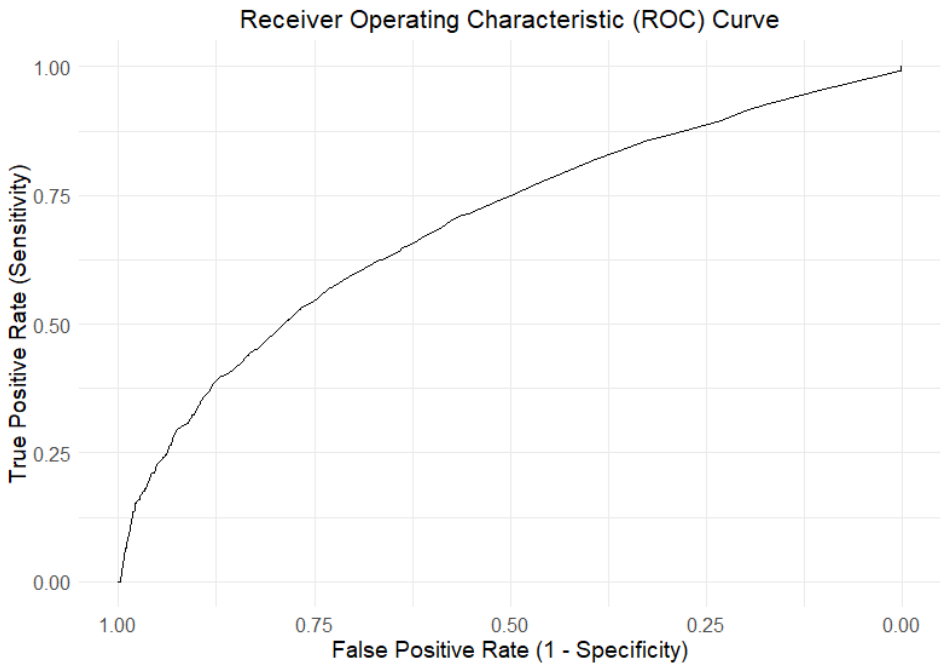

Figure 1: ROC Curve for the Linear Mixed Effects Logistic Regression Model

The coefficients estimated by the LMER model are the following:

- Facebook (baseline): 0
- Instagram: 0.0374
- Online Display: 0.24511
- Online Video: -0.12356
- Paid Search: 0.33529

Since these coefficients are in the log-odds format, they are first exponentiated in order to obtain the odds ratio instead, as this is more interpretable. As such, in the case of Instagram, the odds ratio is circa 1.031. This value is then compared to the baseline channel, which is Facebook. Hence, this value signifies that the odds of a customer converting are around 3.1% higher when they use Instagram, as opposed to Facebook. As for Online Display, the odds ratio becomes around 1.278, indicating that conversion odds increase by approximately 27.8% when using Online Display over Facebook. The odds ratio of Online Video is around 0.884, implying that the odds of converting decrease by circa 11.6% when using Online Video over Facebook. Finally, the odds ratio of the Paid Search channel is 1.399. This suggests that, when compared to Facebook, the use of Paid Search increases the odds of conversion by about 39.9%.

Finally, utilizing these values, one can obtain the credit allocation percentages suggested by the LMER model for these channels. Here, Paid Search obtained the highest score, with 25.0% of conversions being credited to this channel. Online Display attains second place with 22.9%. This is followed by Instagram with 18.4%, Facebook with 17.9%, and lastly, Online Video with 15.8%.

6.4 Gaussian Process Boosting Algorithm

The GPboost model predicted the correct class circa 80.0% of all instances. In finer detail, the precision of the model was 0.754, meaning that circa 75.4% of the predictions of the positive class were accurate. The recall, or sensitivity, obtained was 0.883; the model captures approximately 88.34% of actual positive instances. The F1 score stands at 0.814. At last, the Area Under the ROC Curve (AUC) of the GPBoost model is 0.859. The respective ROC curve is displayed in the figure below:

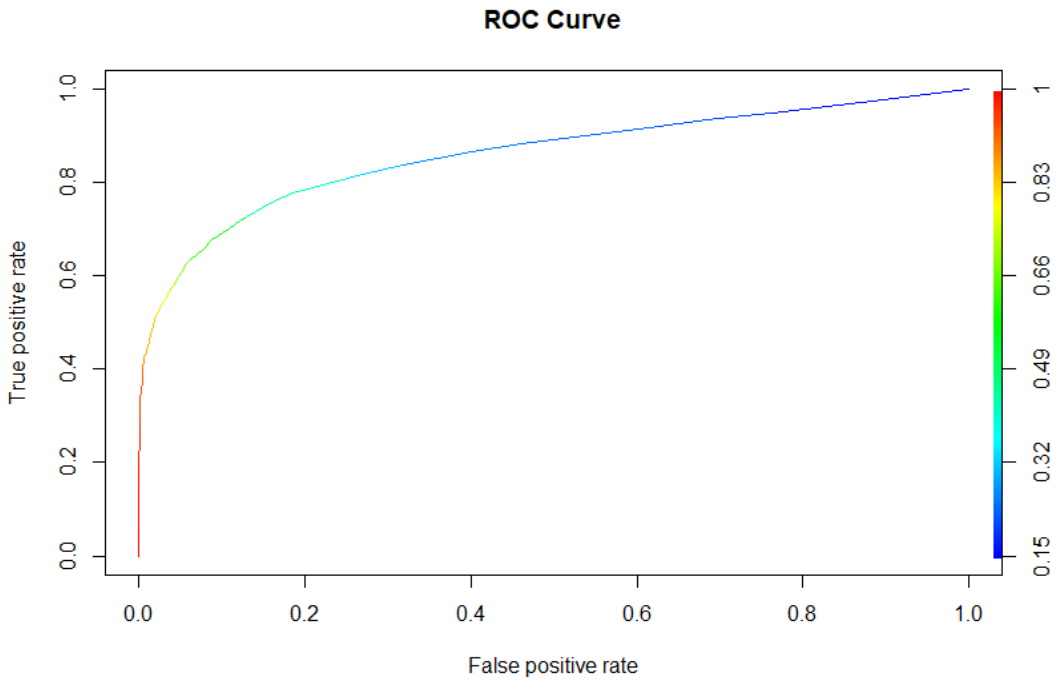

Figure 2: ROC Curve for the Gaussian Process Boosting Model

Following this, a Shapley Value analysis was employed to decipher the feature importance and channel contributions in the GPBoost model. The first channel with the highest Shapley Value

of 0.153 is the Paid Search channel. This indicates its relatively prominent role in shaping the predictions. The distinct feature value distribution, including a dominant negative effect and less pronounced positive influences, suggest that there are varied impacts of individual features within this channel. The subsequent Online Display channel obtained a Shapley Value of 0.084, also signifying its relative impact on the prediction outcomes. Nonetheless, its feature values, particularly a dominant negative effect alongside relatively minor positive contributions, similarly collectively underscore its influential dynamics. Online Video attains a Shapley Value of 0.052, showcasing a medium relative contribution to the outcome. A pivotal positive influence, coupled with more subtle opposing effects, reflects the nuanced interplay of features within this channel. Facebook acquired a Shapley Value of 0.006, still exerting some minimal influence on the outcome. The nearly neutral high feature value and minimal yet balanced low feature values drive this subtle contribution. At last, the Instagram channel received a Shapley Value of 0.002, encapsulating its limited impact. A near-neutral high feature value, paired with correspondingly low, balanced feature values, illustrates its minor role in the predictive landscape. The corresponding Shapley Value Figure is shown below:

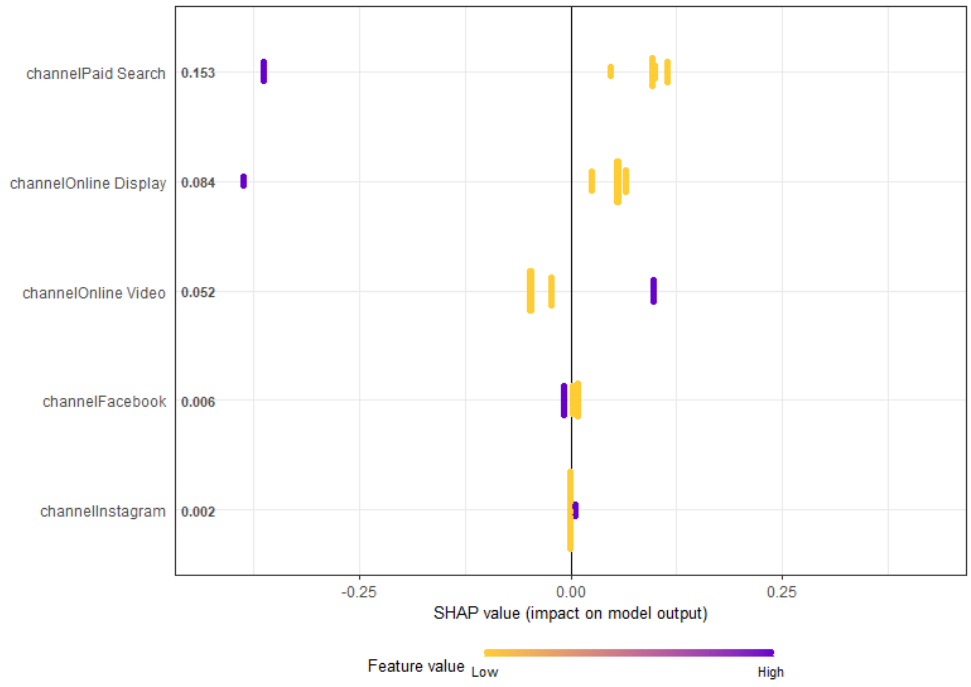

Figure 3: Plot of the Shapley Values for the GPBoost Model

By making use of these Shapley Values, one can determine the respective credit allocation for the various marketing channels. With this approach, the Paid Search channel attains 51.6% of the attribution credit, followed by Online Display with 28.3%, Online video with 17.5%, Facebook with 2.1%, and Instagram with 0.5%.

For comparative purposes, a regular variable importance plot was also constructed (Figure 4). Akin to Shapley values, the variable importance (VI) measure quantifies the contribution of individual variables in the GPBoost model's predictions. Thus, it provides insights into the impact each feature, or channel in this case, had on the model's performance and output. The distribution of channels and their respective importance, as depicted in the VI plot, closely aligns with the findings from the Shapley value analysis. Paid Search emerges as the most significant channel, trailed by Online Display, Online Video, with Facebook and Instagram occupying the subsequent positions in descending order of importance.

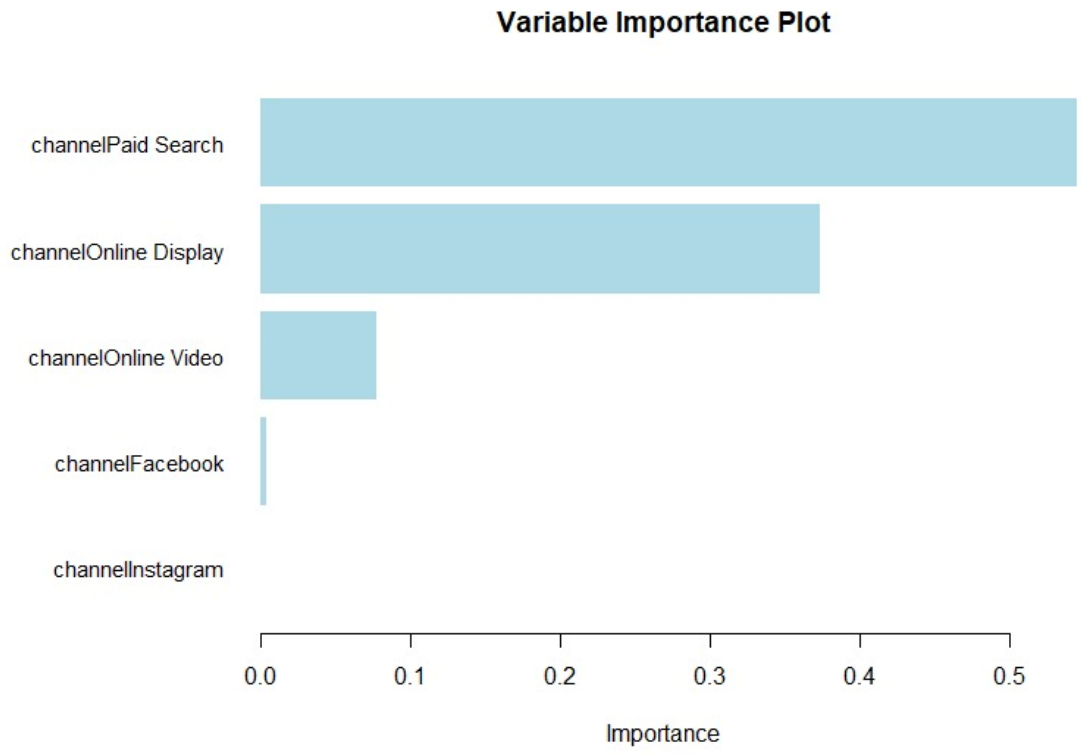

Figure 4: Variable Importance Plot for the GPBoost Model

	Last Touch Attribution	Uniform Attribution	LMER	GPBoost
Facebook	30.1%	29.6%	17.9%	2.1%
Instagram	12.7%	12.8%	18.4%	0.5%
Online Display	12.1%	12.0%	22.9%	28.3%
Online Video	19.3%	19.0%	15.8%	17.5%
Paid search	25.8%	26.5%	25.0%	51.6%

Table 1: A summary of the various % credit attribution values created by the different models.

7 Conclusion and Discussion

The field of attribution modeling stands at a confluence where traditional heuristic approaches meet advanced machine learning methodologies. This section endeavors to dissect and discuss the results obtained in the section prior, with the aim to provide more actionable insights.

Evaluating the accuracy of credit attribution in heuristic models is complex, yet the similarity in attribution percentages highlights their intrinsic reliability. Additionally, most customer journeys were short, featuring single-channel use followed by gradual engagement with more channels. Studies have affirmed the effectiveness of these heuristics in such scenarios. However, it's important to note that these models overlook intricate data dynamics, potentially rendering them less suitable when multiple channels are involved. Therefore, the machine learning models were employed.

The first machine learning model, the linear mixed effects logistic regression, provides a white box, interpretable manner to predict customer conversions, whilst taking into account the random effects present in the data. Seeing as the model attained an AUC of circa 0.70, the overall performance of the model at predicting customer conversions is average. Nonetheless, this average outcome is balanced with ease in interpreting its outcomes and obtaining the credit distributions. Notably, this model resulted in similar allocations for the Paid Search channel as the heuristic approaches did. Surprisingly, the Online Display and Instagram channels, which registered as having the lowest and second lowest values in both heuristics, obtained the second and third highest credit allocations in the linear mixed effects logistic regression (LMER) model.

In contrast, Facebook and Online Video garnered relatively lower allocations in comparison to the heuristic models.

The linear mixed effects logistic regression model uniquely accounts for intricate dependencies between channels used by customers, which may explain the altered significance of the channels. Recognizing these dependencies is crucial, as it distances itself from the isolationist perspective, where channels operate independently. In reality, a customer's interaction with one channel might magnify or diminish the impact of a subsequent channel in their journey. For instance, the increased significance of Online Display and Instagram in the LMER model, as compared to their heuristic rankings, might suggest that these channels are pivotal in particular sequences. They may act as 'catalysts', making subsequent channels more effective, or even amplifying the resonance of preceding channels. Conversely, the decreased significance of Facebook and Online Video suggests that when viewed within the broader ecosystem of channels — and not in isolation — their individual contribution to conversions might be more nuanced, or possibly diluted.

The GPBoost model proved to be performant, attaining an AUC of circa 0.86. This implies that the model was able to effectively classify customer conversions. This success can be attributed to GPBoost's unique capability to harness the predictive strength of XGBoost while also considering the intricate interdependencies among observations associated with the same cookie. Noteworthy is the model's alignment with the LMER approach in attributing the lion's share of credit to the Paid Search channel, impressively accounting for 51.6% of the credit allocation. Similarly, paralleling the LMER model's outcomes, the subsequent channel was Online Display, while Facebook followed as the fourth most impactful channel. The coherence observed across these models underscores that the intrinsic random effects and dependencies captured within them could be instrumental in elevating the significance of these specific channels. Surprisingly, Facebook, attained a 2.1% credit allocation for the GPBoost-Shapley Value approach, making it the second-lowest scored channel. This strongly challenges the values obtained by the heuristic approaches, where Facebook attained the highest credit attribution. Lastly, the GPBoost model allocated a measly 0.5% to the Instagram channel, which is also drastically different from all the other credit allocations from the other models.

8 Limitations of this Research

The pursuit of sophisticated attribution models, while promising, reveals several limitations that must be acknowledged. To begin with, the machine learning models employed require not

only specialized skills in data science but also in attribution modeling. Seeing as the combination of these skills is likely very niche, the practicality for widespread adoption of these data-driven approaches is currently unrealistic. This challenge is exacerbated by the fact that most businesses lack data in the necessary format and/or volumes required for these advanced methodologies. Smaller businesses face the challenge of model implementation due to limited resources, while even larger enterprises encounter obstacles due to the specialized nature of the models and the need for meticulous data preparation. Nonetheless, the current landscape, dominated by LTA, reinforces the need for robust models, particularly when dealing with longer customer journeys. Simultaneously, the field of attribution modeling ought to make advances in creating models that are not only reproducible and standardizable but also more accessible and easier to use by a larger audience. To be effective, these models need to be more user-friendly and comprehensible for businesses, which might not have specialized expertise in attribution modeling.

Another issue faced by this field and research is the reliance on cookie-based data, which faces increasing regulatory constraints. Furthermore, a new session may begin after a certain elapsed time and customers increasingly switch between devices; hence, cookies may not provide a complete view of customer journeys. In addition, the origin of the dataset used in this research remains ambiguous, leaving uncertainty regarding its authenticity as real-world data or a purposely generated dataset for model evaluation. Despite this uncertainty, the dataset was employed in this study as a result of the absence of a more suitable alternative. Consequently, the behaviors manifested within this dataset might not dependably mirror real consumer behaviors.

The application of the GPBoost model, while exhibiting superior predictive performance, introduces complexities. The model evaluations and predictions demonstrate performant outcomes. That being said, the interpretability of the GPBoost model, and the reliability of the Shapley values obtained, warrants careful consideration, potentially diminishing its utility in providing actionable insights. The novelty of the GPBoost package, combined with limited research on its integration with Shapley value analysis in attribution modeling specifically, may contribute to the unusual values observed in the Shapley value analysis.

Finally, it is fundamentally challenging to determine which split for credit allocation between the channels produced by the various models is best. As such, the evaluation of these credit allocations is difficult to assess. A comprehensive assessment would ideally involve a marketing specialist well-versed in each channel's intricacies and market-specific dynamics.

All in all, as the field evolves, refining attribution methods to strike a balance between complexity and practicality remains a central challenge in harnessing the full potential of attribution modeling.

9 Relevance of This Research

This research aims to bridge a gap in the field of attribution modeling by investigating the utilization of XGBoost in conjunction with Shapley values to quantify the impact of various touchpoints on customer conversions. The relevance of this study lies in the scarcity of literature on the application of XGBoost in attribution modeling, especially when combined with Shapley values. In addition, by conducting a comprehensive comparative analysis between the XGBoost model with Shapley value analysis and LTA, this study seeks to provide valuable insights into the relative performance and effectiveness of these approaches. This knowledge will facilitate further research and exploration of novel methodologies in the pursuit of more accurate and reliable attribution modeling techniques. As such, the overall goal of this research is to both fill an existing gap in the current research on attribution modeling and to encourage firms to make more educated decisions when it comes to touchpoint credit allocation by highlighting the potential benefits of using a data-driven approach.

This research equips businesses with evidence-based insights for selecting the appropriate attribution model based on specific marketing channels and targets. Businesses can utilize this information to optimize their resource allocation, directing investments towards impactful touchpoints and maximizing return on investment. Additionally, the use of Shapley values with the XGBoost model enhances the interpretability and transparency of the attribution model. It provides a deeper understanding of each touchpoint's contribution to conversions, which further facilitates efforts in fine-tuning marketing strategies and making informed budget/resource allocation-related decisions. By focusing on channels that drive the highest conversions, businesses can optimize their marketing efforts and achieve better results. Furthermore, this research provides valuable insights into the role of touchpoints in customer decision-making. By implementing the proposed model, the relative contribution of different channels to customer conversion can be uncovered, allowing businesses to tailor their strategies, gain a competitive edge, and improve overall performance and revenue.

10 Bibliography

Abhishek, V., Fader, P. S., & Hosanagar, K. (2012). Media exposure through the funnel: A model of multi-stage attribution. Available at SSRN 2158421.

Anderl, E., Becker, I., von Wangenheim, F., & Schumann, J. H. (2016). Mapping the customer journey: lessons learned from graph-based online attribution modeling. *International Journal of Research in Marketing*, *33*(3), 457–474. https://doi.org/10.1016/j.ijresmar.2016.03.001

Balkanski, E., & Singer, Y. (2015). Mechanisms for Fair Attribution. In Proceedings of the Sixteenth ACM Conference on Economics and Computation (pp. 529-546). Retrieved from https://doi.org/10.1145/2764468.2764505

Berman, R. (2018). Beyond the Last Touch: Attribution in Online Advertising. Foundations and Trends® in Marketing, 12(4), 223-283. https://doi.org/10.1287/mksc.2018.1104

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Buhalis, D., & Volchek, K. (2021). Bridging marketing theory and big data analytics: the taxonomy of marketing attribution. *International Journal of Information Management*, *56*. https://doi.org/10.1016/j.ijinfomgt.2020.102253

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM.

Cui, T. H., Ghose, A., Halaburda, H., Iyengar, R., Pauwels, K., Sriram, S., Tucker, C., & Venkataraman, S. (2021). Informational challenges in omnichannel marketing: Remedies and future research. Journal of Marketing, 85(1), 103-120.

Dalessandro, B., Stitelman, O., & Perlich, C. (2015). Causally motivated attribution for online advertising. In Proceedings of the 24th ACM international on conference on information and knowledge management (pp. 1555-1564).

de Haan, E., Wiesel, T., & Pauwels, K. (2016). The effectiveness of different forms of online advertising for purchase conversion in a multiple-channel attribution framework. International Journal of Research in Marketing, 33(3), 491-507. https://doi.org/10.1016/j.ijresmar.2015.12.001

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Google Marketing Platform. (2023a). *Enterprise Advertising & Analytics Solutions*. Google Marketing Platform. https://marketingplatform.google.com/about/enterprise/

Google Marketing Platform. (2023b). *Free business analytics solutions*. Google Marketing Platform. https://marketingplatform.google.com/about/small-business/

Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. Journal of Big Data, 5(1), 1-13. https://doi.org/10.48550/arXiv.1802.03396

Huang, X., & Marques-Silva, J. (2023). The Inadequacy of Shapley Values for Explainability. arXiv preprint arXiv:2302.08160. Retrieved from https://arxiv.org/abs/2302.08160

Kadyrov, T., & Ignatov, D. I. (2019). Attribution of customers' actions based on machine learning approach.

Kannan, P. K., Reinartz, W., & Verhoef, P. C. (2016). The path to purchase and attribution modeling: Introduction to special section. International Journal of Research in Marketing, 33(3), 425-426. https://doi.org/10.1016/j.ijresmar.2016.07.001

Kireyev, P., Pauwels, K., & Gupta, S. (2016). Do display ads influence search? Attribution and dynamics in online advertising. International Journal of Research in Marketing, 33(3), 475-490. https://doi.org/10.1016/j.ijresmar.2015.09.007

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pages 4765–4774.

Lundberg, S. M. and Lee, S.-I. (2019). Tree shap: Efficient explanations for tree ensemble models. In Proceedings of the 35th International Conference on Machine Learning (ICML), pages 4495–4504.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. arXiv preprint arXiv:1802.03888.

Mahboobi, S.H., Usta, M., & Bagheri, S.R. (2018). Coalition Game Theory In Attribution Modeling: Measuring What Matters at Scale. Journal of Advertising Research, DOI: 10.2501/JAR-2018-014.

Nisar, T., & Yeung, M. (2015). Purchase conversions and attribution modeling in online advertising: An empirical investigation. In 44th EMAC Annual Conference - Collaboration in Research (pp. 1-8). Leuven, Belgium: EMAC.

Shao, X., & Li, L. (2011). Data-driven multi-touch attribution models. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11 (pp. 1054-1062). doi:10.1145/2020408.2020453.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the theory of games (Vol. II, pp. 307-317). Princeton University Press.

Sharma, A., Li, H., & Jiao, J. (2022). The Counterfactual-Shapley Value: Attributing Change in System Metrics. arXiv preprint arXiv:2208.08399 [cs.LG]. Retrieved from https://arxiv.org/abs/2208.08399.

Sigrist, F. (2023). Mixed Effects Machine Learning for High-Cardinality Categorical Variables — Part II: GPBoost library. A demo of GPBoost in Python & R using real-world data. Towards Data Science. https://towardsdatascience.com/mixed-effects-machine-learning-for-high-cardinality-categorical-variables-part-ii-gpboost-3bdd9ef74492

Singal, R., Besbes, O., Desir, A., Goyal, V., & Iyengar, G. (2022). Shapley meets uniform: an axiomatic framework for attribution in online advertising. Management Science, (20220121). https://doi.org/10.1287/mnsc.2021.4263

Starmer, J. [StatQuest with Josh Starmer]. (2020, February 10). XGBoost Part 3 (of 4): Mathematical Details [Video]. YouTube. https://www.youtube.com/watch?v=ZVFeW798-2I&t=147s&ab_channel=StatQuestwithJoshStarmer

Thakurani, S. (2022). Comparison of Channel Attribution Methods to Help Marketers Optimize Marketing Spend. International Research Journal of Modernization in Engineering Technology and Science, 4(2), 293.

W3Techs. (2023, March). Usage of Google Analytics for websites. Retrieved March 25, 2023, from https://w3techs.com/technologies/details/ta-googleanalytics/all/all

Zhao, K., Mahboobi, S. H., & Bagheri, S. R. (2018). Shapley Value Methods for Attribution Modeling in Online Advertising. Retrieved from arXiv website: https://doi.org/10.48550/arXiv.1804.05327